论大数据代数(BDA):大数据科学与工程的分析方法

WANG Yingxu[1,2,3],靳瑾[2]

1. 清华大学大数据系统软件国家工程重点实验室,北京 100084
2. 清华大学北京信息科学与技术国家研究中心,北京 100084
3. 国际认知信息学与认知计算学会;卡尔加里大学电气与计算机工程系,Schulich工程学院,Hotchkiss脑科学研究所, 加拿大卡尔加里 T2N 1N4

摘要　大数据科学和系统的基础研究推动了大数据数学理论的产生。本文引出大数据科学和工程的一种严格分析方法:大数据代数(BDA)从提取各种大数据系统的共同模式中形式地导出大数据科学的数学模型。BDA揭示了任何大数据系统是一种超越传统纯数字的新型递归类型化超结构(RTHS)。基于大数据的递归超结构,创建了一组严格的代数算子,用于对大数据系统的建模、分析、综合和认知学习。这一基础研究建立了一个大数据科学的理论架构,其为解释大数据的原理和性质及其在大数据工程中的形式推理提供了一个方法论基础。

关键词　大数据;数学模型;递归超结构;代数;算法

大数据是人类社会信息时代的代表性现象之一[1-21]。数据科学和信息科学用于量化的数域曾经不断扩展,经历了从二进制数($\mathbb{B}$)、自然数($\mathbb{N}$)、整数($\mathbb{Z}$)、实数($\mathbb{R}$)、复数($\mathbb{C}$)、模糊数($\mathbb{F}$),到超结构数($\mathbb{H}$,简称超数)的演化[17,22-24]。超数$\mathbb{H}$是数域在大数据时代跨越计算机科学、信息科学、认知信息学、计算智能、云计算和社交网络的最新扩展[22,25]。大数据科学的基础研究在近期有两个基础发现[19]:(1) 大数据的数域已越出了传统的$\mathbb{R}$域和与其相适用的数学分析工具的能力;(2) 基于$\mathbb{H}$的大数据概念模型可表达为一个递归类型化n-元组。这些基础研究揭示了大数据科学和工程对传统基于数字计算的理论和方法的挑战以困难所在。

定义1　大数据(big data)是有别于无类型离散数字的超大规模异构类型化超结构(hyperstructure)[22,25],其复杂性涵盖数据的量和质、多样性、存

收稿日期:2019-11-09; 修回日期:2020-01-19
基金项目:国家重点研发计划项目(2016YFB0501504);国家自然科学基金项目(U1509213)
作者简介:WANG Yingxu,教授,研究方向为认知信息学、软件科学、大数据代数和指称数学,电子信箱:yingxu@ucalgary.ca
引用格式:WANG Yingxu, 靳瑾. 论大数据代数(BDA):大数据科学与工程的分析方法[J]. 科技导报, 2020, 38(3): 47-67; doi: 10.3981/j.issn.1000-7857.2020.03.003

储、检索、提取、计算、语义认知、维护和处理。

本文探讨一般大数据系统及其处理的数学基础、基本方法和大数据代数的形式理论。大数据的基本特征是非结构化、异构、单调增长、非描述性、混合/模糊语义,且一致性随时间衰减[19,26]。由于这些固有的复杂性和极大规模的多维超结构对象,不但大数据处理的各个方面和阶段均面临前所未有的问题,而且大数据科学的理论和大数据工程的方法也面临全新的挑战。

1 大数据系统的数学模型与超结构体系

大数据是广泛起源于科学、工程、经济和社会领域大规模异构数据的复杂系统[6,27-32]。本节在类型理论和递归超结构的基础上建立一组大数据系统的数学模型。

1.1 大数据的超结构论域

大数据科学的数学结构的整体性质和公理可以通过对大数据系统的抽象模型归纳而得出。这一形式方法从研究抽象数据和大数据系统的论域开始。

定义2 大数据的论域(the universe of discourse),U是一个6-元组

$$U\triangleq(E, T, Q, R, V, H) \tag{1}$$

式中,E为一实体和/或其可测量属性的有限集;T为E的类型或性质集;Q为E上的一量化标度集;R为一量化关系集($Q\times E$)或限定关系集($E\times T$);V为一类型化的值集$Q\times E\times T\to V|$,其中$|$ 表示一个类型后缀;H为一超结构集$E\times V\times T$。

论域U揭示一般大数据比任何传统数据更加复杂、多维、异构和结构化。在U的基础上,将大数据系统的数学模型将被形式化引出,分为基本(2维)和一般(n维)大数据模型。

1.2 大数据的类型理论

从计算的角度来看,数据的一个重要抽象属性是它们的类(type)或形态。在类型理论中[22, 33-39],类决定了某一形态数据的域、单位和被允许的操作。一个类型系统指定了所有数据对象类的建模和操作规则[18, 35, 38-39]。在一给定完备类型系统中,任何数据对象都可以被指定到一个类型或一个有限类型集的约束。每种类都是一个集合,其中所有数据对象共享一组公共属性、域约束和预定义操作。类可以分为基本和复杂类型。前者是一组最简原始类型,其不能被进一步约减;而后者是由基本类型按照一定的类规则组合而成的混合类型。

定义3 数据对象O是一个有指定类$\mathbb{T}$的变量v, $v\in V\subset$P ⊏ ,其由类型约束$\mu_c(\mathbb{T})$限制以便将此一般类$\mathbb{T}$通过子域裁剪$\mathbb{T}''$转化为一给定问题的特定类$\mathbb{T}'$,即

$$O\triangleq<v:\mathbb{T}\,|\,\mu_c(\mathbb{T})=\mathbb{T}\backslash\mathbb{T}''=\mathbb{T}'> \tag{2}$$

式中,PV表示V的一个幂集;$\mu_c(\mathbb{T})=\mathbb{T}\backslash\mathbb{T}''=\mathbb{T}'$将一般数域$V$中的$\mathbb{T}$裁剪成问题域$V_p$中的子集$\mathbb{T}'$,即$v|\mathbb{T}'=v|\mathbb{T}\backslash v|\mathbb{T}''$。

定义4 大数据的类型系统 T 是一组基本类型T_p和复杂类型T_c的集合[25],即

$$\begin{aligned}\mathsf{T}&\triangleq\mathsf{T}_p\cup\mathsf{T}_c\\&=\{\mathbb{N},\ \mathbb{Z},\ \mathbb{R},\ \mathbb{S},\ \mathbb{L},\ \mathbb{B},\ \mathbb{H}x,\ \mathbb{P}\}\\&\cup\{\mathbb{T},\ \mathrm{SM},\ \mathrm{TX},\ \mathrm{A},\ \mathrm{F},\ \mathrm{V},\ \mathrm{T},\ \mathrm{D},\ \mathrm{TM},\ \mathrm{Bir}\}\end{aligned} \tag{3}$$

式中的基本数据类型T_p分别表示诸如自然数($\mathbb{N}$)、整数($\mathbb{Z}$)、实数($\mathbb{R}$)、字符串($\mathbb{R}$)、逻辑变量($\mathbb{L}$)、字节($\mathbb{B}$)、十六进制数($\mathbb{H}x$)和指针($\mathbb{P}$)类型。复杂数据类型T_c包括诸如任意类($\mathbb{T}$)、结构模型(SM)、文本(TX)、音频(A)、照片(F)、视频(V)、时间(T)、日期(D)、日期-时间(TM)和知识(Bir)。

一般认为,大数据是一种可以通过类型化元组进行形式化描述的超结构[31]。

定义5 数据和数量的超结构类型是一个类型化的n元组,$\tau^n|$,其在满足一定约束条件$\mathbb{T}_i'$下,将n个数据对象封装在异构类$\mathbb{T}_i$ ($1\leqslant i\leqslant n$)的超结构中,即

$$\tau^n|\ \triangleq\left(\mathop{R}_{i=1}^{n}O_i|\mathrm{SM}\right)=\left(\mathop{R}_{i=1}^{n}<v_i:\mathbb{T}_i'|\,\mathbb{T}_i'=\mathbb{T}_i\backslash\mathbb{T}_i''>\right) \tag{4}$$

式中,$\mathop{R}_{i=1}^{n}S_i$表示一个重复结构或迭代行为[40];$\mathbb{T}_i'$是对标准类型$\mathbb{T}_i$的一个类型约束。

数域的类型约束广泛用于指定用户定义或问题相关的类型。其是大数据建模中用于构造或细化数据类型的一般方法。

1.3　基本大数据系统的数学模型

定义6　基本大数据系统(BDS)的数学模型 $\Theta^2=\overset{n}{\underset{i=0}{R}}\overset{m}{\underset{j=0}{R}}d_{ij}|\mathbb{T}_{0j}$ 是论域 U 中的一个二维 $n\times m$ 类型化超结构,其中数据元素 $d_{ij}|\mathbb{T}_{0j}$ 由类型后缀 $|\mathbb{T}$ 指定,可以是任意一种定义在 T 中的基本或复杂类型,即

$$\Theta^2\triangleq\overset{n}{\underset{i=0}{R}}\overset{m}{\underset{j=0}{R}}d_{ij}|\mathbb{T}_{0j}=\begin{bmatrix}\theta_0 & e_1|\mathbb{T}_{01} & e_2|\mathbb{T}_{02} & \cdots & e_m|\mathbb{T}_{0m}\\ \hline \kappa_1 & d_{11}|\mathbb{T}_{01} & d_{12}|\mathbb{T}_{02} & \cdots & d_{1m}|\mathbb{T}_{0m}\\ \kappa_2 & d_{21}|\mathbb{T}_{01} & d_{22}|\mathbb{T}_{02} & \cdots & d_{2m}|\mathbb{T}_{0m}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_n & d_{n1}|\mathbb{T}_{01} & d_{n2}|\mathbb{T}_{02} & \cdots & d_{nm}|\mathbb{T}_{0m}\end{bmatrix}\tag{5}$$

式中,$\overset{n}{\underset{i=0}{R}}\kappa_i|\mathrm{SM}$ 中的每一行表示一个称作为结构模型 $(|\mathrm{SM})$ 的类型化元组,而 $\overset{m}{\underset{j=0}{R}}e_j|\mathbb{T}_{0j}$ 中的每一列代表一个具有特定类型 $(|\mathbb{T}_{0j})$ 的域[30]。

值得注意的是,在如定义6所示基本大数据数学模型 Θ^2 中的第0行 θ_0 是一个特殊的类型化元组,$\tau^n|\mathrm{SM}$ 称为BDS的模式。模式 θ_0 指定BDS各数据域的结构和约束类。BDS数学模型中采用的类后缀 $|\mathbb{T}$ 可用于定义任意类型的大数据,以便满足容纳广泛异构数据的需求。

例1　给定一个有100万用户和7个数据域的社交网络大数据系统 BDS_1。其二维大数据模型的形式结构 $\Theta^2(BDS_1)$ 可以根据定义6进行严格描述如式(6)所示。其中结构模式 $\theta_0(BDS_1)=(ID|\mathrm{N},\ UName|\mathrm{S},\ GName|\mathrm{S},\ Text|\mathrm{T},\ Voice|\mathrm{A},\ Photo|\mathrm{F},\ Video|V)$ 中的7个域分别代表识别号 $(|\mathrm{N})$、用户名 $(|\mathrm{S})$、组名 $(|\mathrm{S})$、文本数据 $(|\mathrm{T})$、语音数据 $(|\mathrm{A})$、照片数据 $(|\mathrm{F})$、和视频数据 $(|\mathrm{V})$。$\Theta^2(BDS_1)$ 中所示的每一大数据记录都受公共模式 $\theta_0(BDS_1)$ 的约束。

$$\Theta^2(BDS_1)=\overset{1000000}{\underset{i=0}{R}}\overset{7}{\underset{j=0}{R}}d_{ij}|\mathbb{T}_j=\begin{bmatrix}\theta_0(BDS_1) & ID|\mathrm{N} & UName|\mathrm{S} & GName|\mathrm{S} & Text|\mathrm{T} & Voice|\mathrm{A} & Photo|\mathrm{F} & Video|\mathrm{V}\\ \hline \kappa_1 & 0000001 & John & G_{0001} & R_t0000001|\mathrm{T} & R_a0000001|\mathrm{A} & R_p0000001|\mathrm{F} & R_v0000001|\mathrm{V}\\ \kappa_2 & 0000002 & Judy & G_{0301} & R_t0000002|\mathrm{T} & R_a0000002|\mathrm{A} & R_p0000002|\mathrm{F} & R_v0000002|\mathrm{V}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000} & 1000000 & Mike & G_{1806} & R_t1000000|\mathrm{T} & R_a1000000|\mathrm{A} & R_p1000000|\mathrm{F} & R_v1000000|\mathrm{V}\end{bmatrix}\tag{6}$$

定义7　大数据结构 Θ 的关键字域 $\kappa_{i0}(\Theta)|\mathrm{SM}$ 是 Θ 中第0列包括所有自然数索引的一个向量,用以指定每行数据记录的标识符 $\overset{n}{\underset{i=1}{R}}d_{i0}|\mathbb{T}_0$,即

$$\overset{n}{\underset{i=1}{R}}\kappa_i|\mathrm{N}\triangleq\overset{n}{\underset{i=1}{R}}\kappa_{i0}(\Theta|\mathrm{SM})|\mathrm{N}=\overset{n}{\underset{i=1}{R}}d_{i0}|\mathbb{T}_0=(d_{00}|\mathrm{N},\ d_{10}|\mathrm{N},\ d_{20}|\mathrm{N},\ ...,\ d_{n0}|\mathrm{N})\tag{7}$$

基于 Θ 的关键字域模型及其所对应的数据域,可以唯一地确定一个大数据系统的整体结构和数据配置空间。

定义8　大数据系统的结构范式,S_Θ 是一个在其关键字集和数据记录集之间的笛卡尔映射,它决定 $\Theta(BDS)$ 中所有数据对象共享的模式 $\theta_0(BDS)$ 及其类型约束,即

$$S_\Theta\triangleq\overset{n}{\underset{i=1}{R}}\kappa_{i0}|\mathrm{SM}\times\overset{m}{\underset{j=1}{R}}\theta_{0j}|\mathrm{SM}\tag{8}$$

例2　如例1所给大数据系统 $\Theta^2(BDS_1)$ 的结构范式 S_Θ 可由定义8确定为

$$S_\Theta(BDS_1)\triangleq\overset{n}{\underset{i=1}{R}}\kappa_{i0}|\mathrm{SM}\times\overset{m}{\underset{j=1}{R}}\theta_{0j}|\mathrm{SM}=\overset{1000000}{\underset{i=1}{R}}\kappa_i|\mathrm{N}\times\{ID_i|\mathrm{N},\ UName|\mathrm{S},$$

$$GName|\mathrm{S}, Text|\mathrm{TX}, Voice|\mathrm{A}, Photo|\mathrm{F}, Video|\mathrm{V}\}$$

引理1 在U上一个基本大数据系统$\Theta(BDS)$中的任何数据对象$d_{ij}|\mathbb{T}_{0j}$的定位,可以由一对索引$(i|\mathrm{N}, j|\mathrm{N})$所表示的一个指针精确地确定,即

$$\rho_e\left(d_{ij}|\mathbb{T}_{0j}\right) \triangleq \Theta(i,j)|\mathrm{P} \to d_{ij}|\mathbb{T}_{0j} \tag{9}$$

证明 引理1中的元素定位算子$\rho_e\left(d_{ij}|\mathbb{T}_{0j}\right)$可以根据$\Theta(BDS)$的二维类型化矩阵结构的性质证明,即

$$\begin{aligned}&\forall i|\mathrm{N} \in [1,n] \wedge j|\mathrm{N} \in [1,m] \wedge d_{ij}|\mathbb{T}_{0j} \sqsubset \Theta,\\ &\rho_e\left(d_{ij}|\mathbb{T}_{0j}\right)|\mathrm{P} = addr_\Theta\left(d_{ij}|\mathbb{T}_{0j}\right)|\mathrm{P}\\ &\quad = \Theta(i,j)|\mathrm{P} \to d_{ij}|\mathbb{T}_{0j}\end{aligned} \tag{10}$$

推论1 在U上对$\Theta(BDS)$中的复杂数据对象的定位,诸如行定位$\rho_r\left(d_{ij}|\mathbb{T}_{ij}\right)$、列定位$\rho_c\left(d_{ij}|\mathbb{T}_{ij}\right)$、子域定位$\rho_\omega\left(d_{ij}|\mathbb{T}_{ij}\right)$和整体定位$\rho_\Omega\left(d_{ij}|\mathbb{T}_{ij}\right)$,可以通过迭代引用式(9)所给的元素定位算子$\rho_e\left(d_{ij}|\mathbb{T}_{ij}\right)$有效地实现,如式(11)所示。

1) Row allocation:

$$\begin{aligned}\rho_r\left(d_{i_0 j}|\mathbb{T}_{0j}\right) &\triangleq \mathop{R}_{j=1}^{m}\rho_e\left(d_{i_0 j}|\mathbb{T}_{0j}\right)\\ &= \mathop{R}_{j=1}^{m}\left[\Theta(i_0|_{i_0\in[1,n]}, j)|\mathrm{P} \to \Theta(i_0|_{i_0\in[1,n]}, j)|\mathbb{T}_{0j}\right]\end{aligned}$$

2) Column allocation:

$$\begin{aligned}\rho_c\left(d_{i j_0}|\mathbb{T}_{0j}\right) &\triangleq \mathop{R}_{i=1}^{n}\rho_c\left(d_{i j_0}|\mathbb{T}_{0j}\right)\\ &= \mathop{R}_{i=1}^{n}\left[\Theta(i, j_0|_{j_0\in[1,m)})|P \to \Theta(i, j_0|_{j_0\in[1,m]})|\mathbb{T}_{0j}\right]\end{aligned}$$

3) Subrange allocation:

$$\begin{aligned}\rho_\omega\left(d_{i_s j_s}|\mathbb{T}_{0j}\right) &\triangleq \mathop{R}_{i_s=n_1}^{n_2}\mathop{R}_{j_s=m_1}^{m_2}\rho_e\left(d_{i_s j_s}|\mathbb{T}_{0j}\right)\\ &= \mathop{R}_{i_s=n_1}^{n_2}\mathop{R}_{j_s=m_1}^{m_2}\left[\Theta(i_s,j_s)|\mathrm{P} \to \Theta(i_s,j_s)|\mathbb{T}_{0j}\right]\end{aligned}$$

4) Layout allocation:

$$\begin{aligned}\rho_\Omega\left(d_{ij}|\mathbb{T}_{0j}\right) &\triangleq \mathop{R}_{i=1}^{n}\mathop{R}_{j=1}^{m}\rho_e\left(d_{ij}|\mathbb{T}_{0j}\right)\\ &= \mathop{R}_{i=1}^{n}\mathop{R}_{j=1}^{m}\left[\Theta(i,j)|\mathrm{P} \to \Theta(i,j)|\mathbb{T}_{0j}\right]\end{aligned} \tag{11}$$

证明 对U中任给$\Theta(BDS)$,因为引理1所给元素定位算子$\rho_e\left(d_{ij}|\mathbb{T}_{ij}\right)$可以寻址4个范畴内的任一数据单元,所以推论1可以通过迭代地应用引理1逐一证明。

1.4 一般大数据系统的数学模型

在U中如式(5)所定义的基本大数据系统模型Θ^2可扩展到一般的n维类型化超结构大数据系统Θ^n。

定义9 大数据系统的一般数学模型为$\Theta^q = \mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}\cdots\mathop{R}_{i_q=0}^{n_q} d_{i_1 i_2 \ldots i_q}|\mathbb{T}_{i_1 i_2 \ldots i_q}$。该模型在$U$中任何$q$维异构数据可被形式化描述成一个$q$维的递归类型化超结构(RTHS),即

$$\begin{aligned}\Theta^q &\triangleq \mathop{R}_{k=q}^{1}\Theta^k\left(\Theta^{k-1}\right)\\ &= \mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}\cdots\mathop{R}_{i_q=0}^{n_q} d_{i_1 i_2 \ldots i_q}|\mathbb{T}_{i_1 i_2 \ldots i_q}\\ &= \mathop{R}_{k=q}^{1}\begin{bmatrix}\theta_0^k & e_1^k|\mathbb{T}_{01}^k & e_2^k|\mathbb{T}_{02}^k e_2 & \cdots & e_m^k|\mathbb{T}_{0m}^k & \theta_0^{k-1}|\mathrm{P}\\ \kappa_1^k & \tau_{11}^k|\mathbb{T}_{01}^k & \tau_{12}^k|\mathbb{T}_{02}^k & \cdots & \tau_{1m}^k|\mathbb{T}_{0m}^k & \kappa_1^{k-1}|\mathrm{P}\\ \kappa_2^k & \tau_{21}^k|\mathbb{T}_{01}^k & \tau_{22}^k|\mathbb{T}_{02}^k & \cdots & \tau_{2m}^k|\mathbb{T}_{0m}^k & \kappa_2^{k-1}|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_n^k & \tau_{m1}^k|\mathbb{T}_{01}^k & \tau_{m2}^k|\mathbb{T}_{02}^k & \cdots & \tau_{mn}^k|\mathbb{T}_{0m}^k & \kappa_m^{k-1}|\mathrm{P}\end{bmatrix}\end{aligned} \tag{12}$$

式中,$\mathop{R}_{j=1}^{m}\kappa_j^k|\mathrm{SM} = \left(\tau_1^k|\mathbb{T}_{01}^k, \tau_2^k|\mathbb{T}_{01}^k, \ldots, \tau_m^k|\mathbb{T}_{01}^k\right)$表示通过指针$\theta_0^{k-1}|\mathrm{P}$递归链接到大数据系统$\mathop{R}_{i=1}^{n}\kappa_i^{k-1}|\mathrm{SM}$中低层结构的一个类型化元组。

通过对比定义6和定义9,可见二维基本大数据模型Θ^2是一般大数据模型Θ^n的一个特例。一般大数据模型$\Theta^n(BDS)$中的元素除了终端层外都是一个类型元组,而基本大数据模型$\Theta^2(BDS)$中的元素是一个终端数据对象。因此,根据定义9,对于给定$\Theta^0(BDS)$,任何前者均可以通过有限步转变为后者。

例3 在例1的基础上建立一给定社交网络的一般大数据模型$\Theta^3(BDS_2)$,其中$\Theta^0 = \mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{4} d_{ij}|\mathbb{T}_{0j}$。按照定义9,该问题的解是将$\Theta^2(BDS_1)$分层细化为一个3层一般大数据模型,如式(13)所示:

$$\Theta^3(BDS_2)=\mathop{R}_{k=3}^{1}\Theta^k\left(\Theta^{k-1}\right),\ \Theta^0=\mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{4}d_{ij}\left|\mathbb{T}_{0J}\right.$$

$$=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{i_1=0}^{n_1}\mathop{R}_{i_2=0}^{n_2}d_{i_0i_1i_2}\left|\mathbb{T}_{i_0i_1i_2}\right.$$

$$=\Theta^3(BDS_2):\begin{bmatrix}\theta_0^3(BDS_2) & ID|\mathrm{N} & UName|\mathrm{S} & GName|\mathrm{S} & \theta_0^2|\mathrm{P}\\ \kappa_1^3 & 0000001 & John & G_{0001} & \kappa_1^2|\mathrm{P}\\ \kappa_2^3 & 0000002 & Judy & G_{0301} & \kappa_2^2|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^3 & 1000000 & Mike & G_{1806} & \kappa_{1000000}^2|\mathrm{P}\end{bmatrix}$$

$$\Downarrow\Theta^2(BDS_2):\begin{bmatrix}\theta_0^2(BDS_2) & Rec\#|\mathrm{N} & TStamp|\mathrm{TM} & Size|\mathrm{B} & \theta_0^1|\mathrm{P}\\ \kappa_1^2 & 0000001 & TM_1 & S_1 & \kappa_1^1|\mathrm{P}\\ \kappa_2^2 & 0000002 & TM_2 & S_2 & \kappa_2^1|\mathrm{P}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^2 & 1000000 & TM_{1000000} & S_{1000000} & \kappa_{1000000}^1|\mathrm{P}\end{bmatrix}\tag{13}$$

$$\Downarrow\Theta^1(BDS_2):\begin{bmatrix}\theta_0^1(BDS_2) & Text|\mathrm{T} & Voice|\mathrm{A} & Photo|\mathrm{F} & Video|\mathrm{V}\\ \kappa_1^1 & R_t0000001|\mathrm{T} & R_a0000001|\mathrm{A} & R_p0000001|\mathrm{F} & R_v0000001|\mathrm{V}\\ \kappa_2^1 & R_t0000002|\mathrm{T} & R_a0000002|\mathrm{A} & R_p0000002|\mathrm{F} & R_v0000002|\mathrm{V}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \kappa_{1000000}^1 & R_t1000000|\mathrm{T} & R_a1000000|\mathrm{A} & R_p1000000|\mathrm{F} & R_v1000000|\mathrm{V}\end{bmatrix}$$

式中，$\Theta^3(BDS_2)$的每一层自顶向下通过指针$\theta_0^{k-1}|\mathrm{P}$链接到下一层，以使任一$k$层被递归地细化$(\Downarrow)$为一个基本二维结构。在$\Theta^3(BDS_2)$中，其顶层设定为$\Theta_0^3(BDS_2)$，并通过$\theta_0^2|\mathrm{P}$将其链接到下一层的二维结构$\theta_0^2(BDS_2)$。第二层对$\theta_0^2(BDS_2)$进行细化，并通过$\theta_0^1|\mathrm{P}$将其链接到下一层$\theta_0^1(BDS_2)$。然后，第一层将$\Theta^1(BDS_2)$表示为在第0层给出的一组终端数据对象$\Theta^0(BDS_2)=\mathop{R}_{i=0}^{1000000}\mathop{R}_{j=0}^{4}d_{ij}\left|\mathbb{T}_{0j}\right.$。

大数据系统的一般数学模型Θ^n可以用递归层次结构清晰和严格地表达任意n维大数据结构，其中每一层都是一个统一的二维基本子结构Θ^2。例3显示大数据系统Θ^n不再是简单的纯数字一维或二维结构，而是一个递归类型化的超结构(RTHS)，其可分解为一组层次化链接的二维结构Θ^2。Θ^n不但揭示了大数据的RTHS数学实质，而且为现实世界中任何复杂大数据系统的建模和分析提供了一种严格、通用、灵活、高效的层次细化方法。该理论也为大数据在H域中的数学分析奠定了基础。

2　大数据代数的数学框架

在数学、计算和数据工程中已经探讨了一系列对大数据的各种操作，诸如关系、数值、计算、组成、推理、数据库和知识库操作。然而支撑大数据科学的数学理论是指称数学(denotational mathematics)[17]，它不但把经典的逻辑、概率和统计数学手段从实域($\mathbb{R}$)扩展到递归超结构($\mathbb{H}$)，并把大数据工程操作从数值处理扩展到基于数据的信息提取(大数据代数[19])、知识生成(概念代数[41]和语义代数[42])与智能生成(推理代数[2,35])。指称数学为大数据分析中的形式推理提供了一套完整的数学结构和操作方法[43-44]，以实现从大数据到全信息、新知识和高智能的有效转换和升华。

大数据代数提供了一个指称数学框架，用于通过形式化的建模，分析和综合操作及其规则对BDS进行严格地操纵。大数据代数视任何BDS为一个如式(12)所定义的一般大数据系统数学模型的实例。大数据代数将任何复杂BDS的操作转化为一个或一组形式化的代数运算。

定义10 大数据代数BDA是通过严格代数运算操纵形式大数据系统 $\Theta(BDS)$ 的一个指称数学结构，其可表示为 U 中的一个三元组，即

$$BDA \triangleq (\Theta, \bullet, \quad) = (\Theta, (\bullet_m, \bullet_a, \bullet_s), \quad) \tag{14}$$

式中，代数运算符 $\bullet = (\bullet_m, \bullet_a, \bullet_s)$ 代表BDS上的一组建模、分析和综合算子。

定义11 BDA的代数算子 $\bullet_{BDA}$ 是对大数据系统 $\Theta(BDS)$ 的一组形式算子，包含2种建模算子 $\bullet_m$、5种分析算子 $\bullet_a$ 和4种综合算子 $\bullet_h$，即：

$$\bullet_{BDA} \triangleq (\bullet_m, \bullet_a, \bullet_s)$$

$$= \begin{cases} \bullet_m = \{\Omega_\theta(\Theta), \Omega_\emptyset(\Theta)\} & // 建模算子 \\ \bullet_a = \{\alpha(\Theta), \gamma(\Theta), \sigma(\Theta), \tau(\Theta), \delta(\Theta)\} & // 分析算子 \\ \bullet_s = \{\iota(\Theta), \lambda(\Theta), \varpi(\Theta), \psi(\Theta)\} & // 综合算子 \end{cases} \tag{15}$$

式中，建模算子包括模式设定 $\Omega_\theta(\Theta)$ 和BDS初始化 $\Omega_\emptyset(\Theta)$。分析算子包括系统赋值 $\alpha(\Theta)$、检索 $\gamma(\Theta)$、选择 $\sigma(\Theta)$、定时 $\tau(\Theta)$ 和微分 $\delta(\Theta)$。综合算子包括归纳 $\iota(\Theta)$、演绎 $\lambda(\Theta)$、积分 $\varpi(\Theta)$ 和均衡 $\psi(\Theta)$。

BDA的框架结构如图1所示，其11个代数算子按定义11归类在3类操作之中。在图1中，一个BDS的结构由 $\Theta(BDS)|\mathrm{SM}$ 表示，其中类型后缀 $|\mathrm{SM}$ 表示一个超结构模型。BDA提供了一种对复杂大数据集和复杂知识库的严格操纵。它可以作为BDS建模、设计、规范、分析、综合、改进、验证和降低处理复杂性的有效手段。

Big Data Algebra (BDA)

	Input	Operation	Operator		Output
Modeling Operators ($\bullet A$)	$\ominus \rightarrow$	Schema specification	(Θ)	$\rightarrow$	^{0}
	$\Theta^0 \rightarrow$	BDS initialization	$_\emptyset(\Theta)$	$\rightarrow$	$\emptyset$
Analytic Operators ($\bullet A$)	$, d_{ij}\|\mathbb{T}_{0j} \rightarrow$	Assignment	$\sigma(\Theta)$	$\rightarrow$	$'$
	$\rightarrow$	Retrieval	$\gamma(\Theta)$	$\rightarrow$	$d_{ij}\|\mathbb{T}_{0j}$
	$, p\|\mathbb{T}_p \rightarrow$	Selection	$\sigma(\Theta)$	$\rightarrow$	$\Theta(i_0, j_0)\|\mathbb{T}_{0j} \sim p\|\mathbb{T}_p$
	$, t_{ij}\|\mathrm{TM} \rightarrow$	Time stamp	$\tau(\Theta)$	$\rightarrow$	$d_{ij}\|\mathbb{T}_{0j} := t_{ij}\|\mathrm{TM}$
	$\rightarrow$	Differentiation	$\delta(\Theta)$	$\rightarrow$	$\mathrm{d}\,\Theta(i, j_0)/\mathrm{d}i$
Synthetic Operators ($\bullet S$)	$, \underset{i=1,2,r}{R^{r+1}}(s_i \in S \vdash p(s_i)) \rightarrow$	Induction	$\iota(\Theta)$	$\rightarrow$	$\underset{i=1,2,r}{R^{r+1}}(s_i \in S \vdash p(s_i)) \Uparrow \underset{i=l}{R^{\|S\|}} p(s_i)$
	$, a \in S \rightarrow$	Deduction	$\lambda(\Theta)$	$\rightarrow$	$p(S\|\mathbb{T}) \vdash p(a\|\mathbb{T} \mid \mathrm{a} \in \mathrm{S})$
	$\rightarrow$	Integration	$\varpi(\Theta)$	$\rightarrow$	$\int_1^n \Theta(i, j_0)\mathrm{d}i$
	$\rightarrow$	Equilibrium	$\psi(\Theta)$	$\rightarrow$	$\int_1^{\xi^-} \Theta(i, j_0)\mathrm{d}i = \int_{\xi^+}^n \Theta(i, j_0)\mathrm{d}i$

图1 大数据代数(BDA)的体系结构

3 大数据代数的系统建模算子

由定义6、定义9所给出的BDS数学模型揭示了在大数据工程中创建和规范复杂大数据系统的一般形式方法。任何复杂BDS的规范设计和描述可由下列BDA建模算子来实现。基于第2节所建立的BDA的数学框架，本节描述用于BDS体系结构配置操作的一组建模算子。

定义12 在 U 中对BDS的建模操作算子 $\bullet_m$ 是对 $\Theta(BDS)$ 的模式设定 Ω_θ 和初始化 $\Omega_\emptyset$ 操作：

$$\bullet_{m} \triangleq \{\Omega_{\theta}, \Omega_{\varnothing}\} \tag{16}$$

3.1 BDA系统模式设定算子

定义13 在U中对基本$\Theta(BDS)$的模式设定算子$\Omega_{\theta}(\Theta)$,根据式(17)指定所给BDS的架构模式$\theta_0(\Theta)|\mathrm{SM}$,即

$$\begin{aligned}\Omega_{\theta}(\Theta) &\triangleq \theta_0(\Theta)|\mathrm{SM} = \Theta(i_0|_{i_0=0,}\overset{m}{\underset{j=0}{R}}j)|\mathbb{T}_{0j}\\ &=\left(\overset{m}{\underset{j=0}{R}}\left(e_{0j}|\mathbb{T}_{0j}|e_{0j}:=id_{0j}|\mathrm{S}\wedge\mathbb{T}_{0j}^{'}:=\mu_{0j}\left(\mathbb{T}_{0j}\right)\right)\right)\\ &=\left(\theta_0|\mathrm{S}, id_{01}|\mathbb{T}_{01}^{'}, id_{02}|\mathbb{T}_{02}^{'}, ..., id_{0m}|\mathbb{T}_{0m}^{'}\right)\end{aligned} \tag{17}$$

式中,$id_{0j}|\mathbb{T}_{0j}^{'}$表示一指定元组中特定域的标识符,其域范畴受限于由$\mu_{0j}\left(\mathbb{T}_{0j}\right)$表示的与问题相关的类型后缀$\mathbb{T}_{0j}^{'}$。

例4 重用例1中建立的大数据系统$\Theta_1(BDS_1)$,其模式$\theta_0(\Theta_1)|\mathrm{SM}$可根据定义13创建如下:

$$\begin{aligned}\Omega_{\theta}(\Theta_1) &= \theta_0(\Theta_1)|\mathrm{SM}\\ &=\Theta\left(0, \overset{m}{\underset{j=0}{R}}\left(id_{0j}|\mathbb{T}_{0j}^{'}=\mu_{0j}\left(\mathbb{T}_{0j}\right)\right)\right)\\ &=\left(\overset{7}{\underset{j=0}{R}}id_{0j}|\mathbb{T}_{0j}^{'}\right)\\ &=\left(\theta_0(BDS_1)|\mathrm{S}, ID|\mathrm{N}, UName|\mathrm{S}, GName|\mathrm{S},\right.\\ &\quad\left.Text|\mathrm{TX}, Voice|\mathrm{A}, Photo|F, Video|\mathrm{V}\right)\end{aligned}$$

上述对基本BDS的数据对象的模式设定算子可以扩展到一般$BDS\ \Theta^{q}\left(BDS^{q}\right)$上的操作。

定义14 在U中对一般$\Theta^{q}(BDS)$的模式设定算子$\Omega_{\theta}^{q}\left(\Theta^{q}\right)$,可根据式(18)自顶向下分层递归设定该BDS的一组架构模式$\overset{1}{\underset{k=q}{R}}\theta_0^{k}\left(\Theta^{k}\right)|\mathrm{SM}$,即

$$\begin{aligned}\Omega_{\theta}^{q}\left(\Theta^{q}\right) &\triangleq \overset{1}{\underset{k=q}{R}}\theta_0^{k}\left(\Theta^{k}\right)|\mathrm{SM}\\ &=\overset{1}{\underset{k=q}{R}}\left(\overset{m}{\underset{j=0}{R}}\left(e_{0j}|\mathbb{T}_{0j}^{k'}|e_{0j}^{k}:=id_{0j}^{k}|\mathrm{S}\wedge\mathbb{T}_{0j}^{k'}:=\mu_{0j}^{k}\left(\mathbb{T}_{0j}^{k}\right)\right)\right)\\ &=\overset{1}{\underset{k=q}{R}}\left[\left(\theta_0^{k}|\mathrm{S}, id_{01}^{k}|\mathbb{T}_{01}^{k'}, id_{01}^{k}|\mathbb{T}_{01}^{k'}, ..., id_{0m}^{k}|\mathbb{T}_{0m}^{k'}\right)\right],\\ &\quad\mathbb{T}_{0j}^{k'}:=\mu_{0j}^{k}\left(\mathbb{T}_{0j}^{k}\right)\end{aligned} \tag{18}$$

式中,$id_{0j}^{k}|\mathbb{T}_{0j}^{k'}$是第$k$层的第$j$域的给定标识符,其特定类型后缀受$\mu_{0j}^{k}\left(\mathbb{T}_{0j}^{k}\right)$约束。

例5 如例3中给出的大数据社交网络系统BDS_2的形式模型$\mathrm{BDS}_2|\Theta^{3}$可以根据定义14被分层递归创建如下:

$$\begin{aligned}&\Omega_{\theta}^{3}\left(\mathrm{BDS}_2|\Theta_2^{3}\right) \triangleq \overset{1}{\underset{k=3}{R}}\left[\theta_0^{k}\left(\Theta_2^{3}\right)|\mathrm{SM}\right]\\ &=\overset{1}{\underset{k=3}{R}}\left[\left(\theta_0^{k}|\mathrm{S}, id_{01}^{k}|\mathbb{T}_{01}^{k}, id_{01}^{k}|\mathbb{T}_{01}^{k}, ..., id_{0m}^{k}|\mathbb{T}_{0m}^{k}\right)\right]\\ &=\overset{4}{\underset{i_3=0}{R}}\overset{4}{\underset{i_2=0}{R}}\overset{4}{\underset{i_1=0}{R}}e_{i_3i_2i_1}|\mathbb{T}_{i_2i_3i_1}^{k}\\ &=\theta_0^{3}\left(\Theta_2^{3}\right)|\mathrm{SM} = \overset{4}{\underset{i_3=0}{R}}e_{i_3}|\mathbb{T}_{i_3}^{k}\\ &=\left(\theta_0^{3}\left(\Theta_2^{3}\right)|\mathrm{S}, ID|\mathrm{N}, UName|\mathrm{S}, GName|\mathrm{S}, \theta_0^{2}\left(\Theta^{3}\right)|\mathrm{S}\right)\\ &\Downarrow\theta_0^{2}\left(\Theta_2^{3}\right)|\mathrm{SM} = \overset{4}{\underset{i_2=0}{R}}e_{i_2}|\mathbb{T}_{i_2}^{k}\\ &=\theta_0^{2}\left(\Theta_2^{3}\right)|\mathrm{S}, Rec\#|\mathrm{N}, TStamp|\mathrm{TM}, Size|\mathrm{B}, \theta_0^{1}\left(\Theta^{3}\right)|\mathrm{S}\\ &\Downarrow\theta_0^{1}\left(\Theta_2^{3}\right)|\mathrm{SM} = \overset{4}{\underset{i_1=0}{R}}e_{i_1}|\mathbb{T}_{i_1}^{k}\\ &=\left(\theta_0^{1}\left(\Theta_2^{3}\right)|\mathrm{S}, Text|\mathrm{TX}, Voice|\mathrm{A}, Photo|\mathrm{F}, Video|\mathrm{V}\right)\end{aligned}$$

该一般BDS的规范算子表示一个演绎过程$(\Downarrow)$,其中BDS的每一层均可自顶向下指定。

3.2 BDA系统初始化算子

定义15 在U中对基本$\Theta(BDS)$的初始化算子$\Omega_{\varnothing}(\Theta)$,预设$\Theta$为一组给定的初始值$\varnothing_{ij}|\mathbb{T}_{0j}$,其各个域的类型对应于其架构模式$\theta_0(\Theta)|\mathrm{SM}$的规范限定,即

$$\begin{aligned}\Omega_{\varnothing}(\Theta) &\triangleq \overset{n}{\underset{i=1}{R}}\overset{m}{\underset{j=1}{R}}\left(d_{ij}|\mathbb{T}_{0j}:=\varnothing_{ij}|\mathbb{T}_{0j}\right),\ \mathbb{T}_{0j}\in\theta_0(\Theta)|\mathrm{SM}\\ &=\overset{n}{\underset{i=1}{R}}\overset{m}{\underset{j=1}{R}}\Theta(i,j)|\mathbb{T}_{0j}:=\varnothing_{ij}|\mathbb{T}_{0j}\end{aligned} \tag{19}$$

式中每一$\varnothing_{ij}|\mathbb{T}_{0j}$在$\theta_0(\Theta)|\mathrm{SM}$是一对应于给定类型的空值,即$\varnothing_{ij}|\mathbb{T}_{0j}^{'}\in\{\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{S}, \mathbb{L}, \mathbb{B}, \mathbb{H}\mathrm{x}, \mathbb{P}\}=''\{0, 0, 0, '', \mathrm{F}, 0, 0, \varnothing\}$。

例6 基于例3中得到的模式$\theta_0(\Theta_1)|\mathrm{SM}$,大数据系统$\mathrm{BDS}_1$的形式模型$\Theta_1(BDS_1)$可以根据定义15进行初始化如下:

$$\begin{aligned}\Omega_{\varnothing}\left(\Theta_1(BDS_1)\right) &= \overset{1000000}{\underset{i=1}{R}}\overset{7}{\underset{j=1}{R}}\left(d_{ij}|\mathbb{T}_{0j}:=\varnothing_{ij}|\mathbb{T}_{0j}\right),\\ &\quad\mathbb{T}_{0j}\in\theta_0\left(\Theta_1\right)|\mathrm{SM}\\ &\overset{1000000}{\underset{i=1}{R}}\left(d_{i1}|\mathrm{N}:=0, d_{i2}|\mathrm{S}:='', d_{i3}|\mathrm{S}:='', d_{i4}|\mathrm{TX}:='',\right.\end{aligned}$$

$d_{i5}|\mathrm{A}:=(0,0),\ d_{i6}|\mathrm{F}:=(0,0),\ d_{i7}|\mathrm{V}:=(0,0,0))$

其中的复杂数据类型可由式(3)所给基本类型导出,即 $\mathrm{A}|\mathrm{SM}=\mathrm{B}\times\mathrm{T}$, $\mathrm{F}|\mathrm{SM}=\mathrm{B}\times\mathrm{B}$,和 $\mathrm{V}|\mathrm{SM}=\mathrm{B}\times\mathrm{B}\times\mathrm{T}$。

上述对基本BDS的数据对象的初始化算子可以扩展到一般BDS $\Theta^q(BDS^q)$ 上的操作。

定义16 在 U 中对一般 $\Theta^q(BDS)$ 的初始化算子 $\Omega^q_{\varnothing}(\Theta^q)$,设定一组递归的二维结构 $\overset{n}{\underset{i=1}{R}}\ \overset{m}{\underset{j=1}{R}} d^k_{ij}|\mathbb{T}^k_{0j}$ 为对应的初始值 $\overset{n}{\underset{i=1}{R}}\ \overset{m}{\underset{j=1}{R}} \varnothing^k_{0j}|\mathbb{T}^k_{0j}$,其各个域的类型 $|\mathbb{T}^k_{0j}$ 对应于其架构模式 $\theta^k_0(\Theta^k)|\mathrm{SM}$ 在第 k 层的限定,即

$$\begin{aligned}&\Omega^q_{\varnothing}(\Theta^q)\triangleq\overset{1}{\underset{k=q}{R}}\Omega^k_{\varnothing}(\Theta^k)\\&=\overset{1}{\underset{k=q}{R}}\left[\overset{n}{\underset{i=1}{R}}\ \overset{m}{\underset{j=1}{R}}\left(d^k_{ij}|\mathbb{T}^k_{0j}:=\varnothing^k_{0j}|\mathbb{T}^k_{0j}\right),|\mathbb{T}^k_{0j}\in\theta^k_0(\Theta^k)|\mathrm{SM}\right]\\&=\overset{1}{\underset{k=q}{R}}\left[\overset{n}{\underset{i=1}{R}}\ \overset{m}{\underset{j=1}{R}}\Theta^k(i^k,j^k)|\mathbb{T}^k_{0j}:=\varnothing^k_{ij}|\mathbb{T}^k_{0j},\varnothing^k_{ij}\Big|\mathbb{T}^k_{0j}\in\overset{m}{\underset{j=1}{R}}\varnothing^k_{0j}|\mathbb{T}^k_{0j}\right]\end{aligned}\tag{20}$$

式中,$d^k_{ij}|\mathbb{T}^k_{0j}$ 是 Θ^q 中第 k 层二维数据的第 i 行和第 j 列。

例7 基于例5中得到的模式 $\theta_0(\Theta^3_2|\mathrm{SM})$,大数据社交网络系统 BDS_2 的形式模型 $\Theta^3_2(\mathrm{BDS}_2)$ 可以根据定义16被分层初始化如下:

$$\Omega^3_{\emptyset}(\mathrm{BDS}_2|\Theta^3_2)=\overset{1}{\underset{k=3}{R}}\left[\overset{1000000}{\underset{i=1}{R}}\ \overset{|\theta^k_0|}{\underset{j=1}{R}}\left(d^k_{ij}|\mathbb{T}^k_{0j}:=\varnothing^k_{0j}|\mathbb{T}^k_{0j}\right)\right],\ \mathbb{T}^k_{0j}\in\theta^k_0(\Theta^3_2)|\mathrm{SM}$$

$$=\overset{1000000}{\underset{i=1}{R}}\ \overset{4}{\underset{j=1}{R}}\Theta^3_2(BDS_2)\begin{cases}d^3_{i1}|\mathbb{T}_{i1}=ID|\mathrm{N}:=\varnothing|\mathrm{N}=0\\d^3_{i2}|\mathbb{T}_{i2}=UName|\mathrm{S}:=\varnothing|\mathrm{S}=''\\d^3_{i3}|\mathbb{T}_{i3}=GName|\mathrm{S}:=\varnothing|\mathrm{S}=''\\d^3_{i4}|\mathbb{T}_{i4}=Record|\theta^1_2:=\varnothing|\theta^2_0\end{cases}$$

$$\Downarrow\overset{1000000}{\underset{i=1}{R}}\ \overset{4}{\underset{k=1}{R}}\Theta^2_2(BDS_2)\begin{cases}d^2_{i1}|\mathbb{T}_{i1}=Rec\#|\mathrm{N}:=\varnothing|\mathrm{N}=0\\d^2_{i2}|\mathbb{T}_{i2}=TStamp|\mathrm{TM}:=\\\quad\varnothing|\mathrm{YYYY:MM:DD:hh:mm:ss:m}\\d^2_{i3}|\mathbb{T}_{i3}=Size|\mathrm{B}:=\varnothing|\mathrm{B}=0\\d^2_{i4}|\mathbb{T}_{i4}=Content|\theta^2_2:=\varnothing|\theta^1_0\end{cases}$$

$$\Downarrow\overset{1000000}{\underset{i=1}{R}}\ \overset{4}{\underset{p=1}{R}}\Theta^1_2(BDS_2)\begin{cases}d^1_{i1}|\mathbb{T}_{i1}=Text|\mathrm{TX}:=\varnothing|\mathrm{TX}=''\\d^1_{i2}|\mathbb{T}_{i2}=Voise|\mathrm{A}:=\varnothing|\mathrm{A}\\\quad=(0,00{:}00{:}00)\\d^1_{i3}|\mathbb{T}_{i3}=Photo|\mathrm{F}:=\varnothing|\mathrm{F}=(0,0)\\d^1_{i4}|\mathbb{T}_{i4}=Video|\mathrm{V}:=\varnothing|\mathrm{V}\\\quad=(0,0,00{:}00{:}00)\end{cases}$$

4 大数据代数的分析算子

基于前文所建立的BDA的数学框架和范例,本节导出一组对大数据系统的基本分析算子,以实现BDA的演绎操作。任何对BDS的复杂分析操作均可表示为这些基本BDA分析算子的代数组合。

定义17 BDA的分析算子 $\bullet_a$ 是在U中对 $\Theta(BDS)$ 的一组演绎操作,用于对大数据对象在BDS中的定位和分析,即

$$\bullet_a\triangleq\{\alpha(\Theta),\gamma(\Theta),\sigma(\Theta),\tau(\Theta),\delta(\Theta)\}\tag{21}$$

式中,各算子分别代表赋值、检索、选择、定时和微分。

4.1 BDA数据赋值算子

在如引理1和推论1所描述的超结构数据域的基础上,数据对象赋值是向一个在给定层和/或域的大数据对象写入相应类型化数据的操作。

定义18 在 U 中对基本BDS的单一数据元素的赋值算子 $\alpha_e(\Theta)$,将一给定类型的值 $v|\mathbb{T}$ 赋予由一对索引 (i_0,j_0) 所指定的目标数据元素 $d_{i_0,j_0}|\mathbb{T}_{0j}=\Theta(i_0,j_0)|\mathbb{T}_{0j}$,即

$$\begin{aligned}&\alpha_e(\Theta)\triangleq\Theta(i_0,j_0)\big|\mathbb{T}_{0j_0}:=v|\mathbb{T},\\&iff\ \mathbb{T}_{0j_0}=\mathbb{T}\wedge i_0\in[1,n]\wedge j_0\in[1,m]\end{aligned}\tag{22}$$

式22所给单一数据元素的赋值算子可以推广到对BDS的整体或局部大数据系统的赋值。

定义19 在 U 中对基本BDS的数据对象赋值算子 $\alpha_{r/c/\omega}(\Theta)$,可分为行赋值 $\alpha_r\left(\Theta\left(i_0,\overset{m}{\underset{j=1}{R}}j\right)\right)$、列赋值 $\alpha_c\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,j_0\right)\right)$ 和子域赋值 $\alpha_\omega\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}}i,\ \overset{m_k}{\underset{j=1}{R}}j\right)\right)$,其可通过迭代应用式(22)所给单一数据元素的赋值算子

$\alpha_e\left(\Theta(i_0,j_0)\right)$ 实现,即

$$\alpha_{r/c/\Omega}(\Theta)\triangleq \begin{cases} \alpha_r\left(\Theta\left(i_0,\overset{m}{\underset{j=1}{R}}j\right)\right)\triangleq \overset{m}{\underset{j=1}{R}}\alpha_e\left(\Theta(i_0,j)\right) \\ \quad =\overset{m}{\underset{j=1}{R}}\left\{\Theta\left(i_0\big|_{i_0\in[1,n]},j\right)\Big|\mathbb{T}_{0j}:=v_j|\mathbb{T}_j\in\overset{m}{\underset{j=1}{R}}v_j|\mathbb{T}_j\right\} \\ \alpha_c\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,j_0\right)\right)\triangleq \overset{n}{\underset{i=1}{R}}\alpha_e\left(\Theta(i,j_0)\right) \\ \quad =\overset{n}{\underset{i=1}{R}}\left\{\Theta\left(i,j_0\big|_{j_0\in[1,m]}\right)\Big|\mathbb{T}_{0j_0}:=v_i|\mathbb{T}_{j_0}\in\overset{n}{\underset{i=1}{R}}v_j|\mathbb{T}_{j_0}\right\} \\ \alpha_\omega\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}}i,\overset{m_k}{\underset{j=1}{R}}j\right)\right)\triangleq \overset{n_k}{\underset{i=1}{R}}\overset{m_k}{\underset{j=1}{R}}\alpha_e\left(\Theta(i,j)\right) \\ \quad =\overset{n_k}{\underset{i=1}{R}}\overset{m_k}{\underset{j=1}{R}}\left\{\Theta(i,j)|\mathbb{T}_{0j}:=v_{ij}|\mathbb{T}_j\in\overset{n_k}{\underset{i=1}{R}}\overset{m_k}{\underset{j=1}{R}}v_{ij}|\mathbb{T}_j\right\} \end{cases} \tag{23}$$

大数据系统的整体赋值算子等效于将式(23)的子域赋值扩展到各子域的最大值。整体赋值算子也等效于用 $v_{ij}|\mathbb{T}_{0j}$ 取代定义19中所给DBS初始化算子 $\Omega_\varnothing(\Theta)$ 中的 $\varnothing_{ij}|\mathbb{T}_{0j}$。反之,$\Omega_\varnothing(\Theta)$ 亦可通过整体赋值算子来实现,即 $\Omega_\varnothing(\Theta)$ 是当 $v_{ij}|\mathbb{T}_{0j}=\varnothing_{ij}|\mathbb{T}_{0j}$ 时 $\alpha_\Omega\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,\overset{m}{\underset{j=1}{R}}j\right)\right)$ 的一个特例。

例8　设 $i_0=1000000, j_0=2,\ v|\mathbb{T}_2=Mike|\mathrm{S}$,赋值给在例1中所创建的数据对象 $\Theta_1(BDS_1)$ 可以根据定义19操作如下:

$$\begin{aligned}\alpha\left(\Theta_1(BDS_1)\right)&=\alpha_e\left(\Theta_1\left(i_0\big|_{i_0=1000000},j_0\big|_{j_0=2}\right)\right)\\&=\Theta_1(1000000,2)|\mathrm{S}:=v|\mathrm{S}\\&=\Theta_1(1000000,2)|\mathrm{S}:=Mike|\mathrm{S}\end{aligned}$$

上述对基本BDS的数据对象的赋值算子可以扩展到对一般 BDS $\Theta^q\left(BDS^q\right)$ 上的操作。

定义20　在 U 中对一般BDS $\Theta^q\left(BDS^q\right)$ 的数据对象赋值算子 $\alpha^q_{r/c/\omega}\left(\Theta^q\right)$,可分为元素赋值 $\alpha_e\left(\Theta(i,j)\right)$(式(22))、行赋值 $\alpha^k_r\left(\Theta\left(i^k_0,\overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)$、列赋值 $\alpha^k_c\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}}i^k,j^k_0\right)\right)$ 和子域赋值 $\alpha^k_\omega\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}}i^k,\overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)$,其可通过迭代导出到一系 q 层自顶向下的基本操作 $\overset{1}{\underset{k=q}{R}}\alpha^k_{e/r/c/\omega}\left(\Theta^k_q\right)$,即

$$\alpha^q_{e/r/c/\Omega}\left(\Theta^q\right)\triangleq\overset{1}{\underset{k=q}{R}}\alpha^k_{e/r/c/\Omega}\left(\Theta^k_q\right)$$

$$= \begin{cases} \overset{1}{\underset{k=q}{R}}\alpha^k_r\left(\Theta\left(i^k_0,\overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\overset{m_k}{\underset{j^k=1}{R}}\alpha^k_e\left(\Theta\left(i^k_0,j^k\right)\right)\right] \\ \quad =\overset{1}{\underset{k=q}{R}}\overset{m_k}{\underset{j^k=1}{R}}\left[\Theta^k_q\left(i^k_0\big|_{i^k_0\in[1,n_k]},j^k\right)\Big|\mathbb{T}_{0^kj^k}\right. \\ \quad \left.:=v_{i^k_0j^k}|\mathbb{T}_{j^k}\in\left\{\overset{1}{\underset{k=q}{R}}\overset{m_k}{\underset{j^k=1}{R}}v_{kj^k}|\mathbb{T}_{j^k}\right\}\right] \\ \overset{1}{\underset{k=q}{R}}\alpha^k_c\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}}i^k,j^k_0\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\overset{n_k}{\underset{i^k=1}{R}}\alpha^k_e\left(\Theta\left(i^k,j^k_0\right)\right)\right] \\ \quad =\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}\left[\Theta^k_q\left(i^k,j^k_0\big|_{j^k_0\in[1,m_k]}\right)\Big|\mathbb{T}_{0^kj^k_0}\right. \\ \quad \left.:=v_{i^kj^k_0}|\mathbb{T}_{j^k_0}\in\left\{\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}v_{ki^k}|\mathbb{T}_{j^k_0}\right\}\right] \\ \overset{1}{\underset{k=q}{R}}\alpha^k_\omega\left(\left(\Theta\overset{n_k}{\underset{i^k=1}{R}}i^k,\overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\overset{n_k}{\underset{i^k=1}{R}}\overset{m_k}{\underset{j^k=1}{R}}\alpha^k_e\left(\Theta\left(i^k,j^k\right)\right)\right] \\ \quad =\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}\overset{m_k}{\underset{j^k=1}{R}}\left[\Theta^k_q\left(i^k,j^k\right)|\mathbb{T}_{0^kj^k}\right. \\ \quad \left.:=v_{i^kj^k}|\mathbb{T}_{j^k}\in\left\{\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}\overset{m_k}{\underset{j^k=1}{R}}v_{i^kj^k}|\mathbb{T}_{j^k}\right\}\right] \end{cases} \tag{24}$$

当子域赋值的行和列的范围扩展到最大时,其即转化为对一般BDS的整体赋值 $\overset{1}{\underset{k=q}{R}}\alpha^k_\omega\left(\left(\Theta\overset{n_k}{\underset{i^k=1}{R}}i^k,\overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)$。

例9　将一组特定类型值赋予一个如例3所示的一般BDS $\Theta^3_2(BDS_2)$,可以根据定义20逐层迭代操作如下:

$$\alpha^3_\Omega\left(\Theta^3_2(BDS_2)\right)=\overset{1}{\underset{k=3}{R}}\alpha^k_e\left(i^k_0,j^k_0\right)$$

$$=\overset{1}{\underset{k=3}{R}}\Theta^k_2\left(i^k_0\big|_{i^k_0\in[1,n_k]},j^k_0\big|_{j^k_0\in[1,m_k]}\right)\Big|\mathbb{T}_{0j^k_0}:=v_k|\mathbb{T}_k,\text{其中}$$

$$v_3|\mathbb{T}_3=Mike|\mathrm{S},v_2|\mathbb{T}_2=1023|\mathrm{B},\text{and }v_1|\mathbb{T}_1=R_t1000000|\mathrm{TX}$$

$$=\Theta^3_2\left(i^3_0\big|_{i^3_0=1000000},j^3_0\big|_{j^3_0=2}\right)|\mathrm{S}=\Theta^3_3(1000000,2)|\mathrm{S}:=Mike|\mathrm{S}$$

$$\Downarrow\Theta^2_2\left(i^2_0\big|_{i^2_0=1000000},j^2_0\big|_{j^2_0=3}\right)|\mathrm{B}=\Theta^2_3(1000000,3)|\mathrm{B}:=Size|\mathrm{B}$$

$$\Downarrow\Theta^1_2\left(i^1_0\big|_{i^1_0=1000000},j^1_0\big|_{j^1_0=1}\right)|\mathrm{S}=\Theta^1_3(1000000,1)|\mathrm{TX}$$

$$:=R_t1000000|\mathrm{TX}$$

4.2 BDA数据检索算子

根据引理1和推论1,数据对象在BDS中的检索是基于超结构中的数据元素的定位操作所实现的。大数据对象检索算子是数据对象赋值算子的逆操作,用于对BDS的只读提取。

定义21 在U中对基本BDS的数据对象的检索算子$\gamma_{e/r/c/\omega}(\Theta)$,用于由一对索引$(i|\mathrm{N}, j|\mathrm{N})$所指定的数据对象的只读提取操作。其可分为元素检索$\gamma_e\left(\Theta\left(i_0, \overset{m}{\underset{j=1}{R}} j\right)\right)$、行检索$\gamma_r\left(\Theta\left(i_0, \overset{m}{\underset{j=1}{R}} j\right)\right)$、列检索$\gamma_c\left(\Theta\left(\overset{n}{\underset{i=1}{R}} i, j_0\right)\right)$和子域检索$\gamma_\omega\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}} i, \overset{m_k}{\underset{j=1}{R}} j\right)\right)$,即:

$$\gamma_{e/r/c/\omega}(\Theta) \triangleq \begin{cases} \gamma_e\left(\Theta(i_0, j_0)\right) \triangleq v|\mathbb{T}:=\Theta\left(i_0|_{i_0\in[1,n]}, j_0|_{j_0\in[1,m]}\right)|\mathbb{T}_{0j_0} \\ \gamma_r\left(\Theta\left(i_0, \overset{m}{\underset{j=1}{R}} j\right)\right) \triangleq \overset{m}{\underset{j=1}{R}} \gamma_e\left(\Theta(i_0, j)\right) \\ \quad = \overset{m}{\underset{j=1}{R}}\left[v_j|\mathbb{T}_j:=\Theta\left(i_0|_{i_0\in[1,n]}, j\right)|\mathbb{T}_{0j}\right] \\ \gamma_c\left(\Theta\left(\overset{n}{\underset{i=1}{R}} i, j_0\right)\right) \triangleq \overset{n}{\underset{i=1}{R}} \gamma_e\left(\Theta(i, j_0)\right) \\ \quad = \overset{n}{\underset{i=1}{R}}\left[v_i|\mathbb{T}_{j_0}:=\Theta\left(i, j_0|_{j_0\in[1,m]}\right)|\mathbb{T}_{0j_0}\right] \\ \gamma_\omega\left(\Theta\left(\overset{n'}{\underset{i=1}{R}} i, \overset{m'}{\underset{j=1}{R}} j\right)\right) \triangleq \overset{n}{\underset{i=1}{R}}\overset{m}{\underset{j=1}{R}} \gamma_e\left(\Theta(i, j)\right) \\ \quad = \overset{n}{\underset{i=1}{R}}\overset{m}{\underset{j=1}{R}}\left[v_{ij}|\mathbb{T}_j:=\Theta(i, j)|\mathbb{T}_{0j}\right] \end{cases} \tag{25}$$

当子域检索的行和列的范围均为最大时,其即转化为整体检索$\gamma_\Omega\left(\Theta\left(\overset{n}{\underset{i=1}{R}} i, \overset{m}{\underset{j=1}{R}} j\right)\right)$或拷贝。

例10 任给一个基本$\Theta_1(BDS_1)$如例1所示。该大数据对象的元素、行和列检索可以根据定义21实现如下:

$$\begin{aligned} \gamma_e\left(\Theta_1(BDS_1)\right) &= \gamma_e\left(\Theta_1\left(i_0|_{i_0=1000000}, j_0|_{j_0=2}\right)\right) \\ &= v|\mathrm{S}:=\Theta_1(1000000, 2)|\mathrm{S} \\ &= Mike|S \end{aligned}$$

$$\begin{aligned} \gamma_r\left(\Theta_1\left(i_0|_{i_0=2}, \overset{7}{\underset{j=1}{R}} j\right)\right) &= \overset{7}{\underset{j=1}{R}} \gamma_e\left(\Theta_1(2, j)\right) \\ &= \overset{7}{\underset{j=1}{R}}\left(v_j|\mathbb{T}_j:=\Theta_1(2, j)|\mathbb{T}_{0j}\right) \\ &= \big(ID|\mathrm{N}=0000002|\mathrm{N},\ UName|\mathrm{S}=Judy|\mathrm{S}, \\ &\quad GName|\mathrm{S}=G_{0301}|\mathrm{S},\ Text|\mathrm{TX}=R_t\ 0000002|\mathrm{TX}, \\ &\quad Voice|\mathrm{A}=R_a\ 0000002|\mathrm{A},\ Photo|\mathrm{F}=R_p\ 0000002|\mathrm{F}, \\ &\quad Video|\mathrm{V}=R_v\ 0000002|\mathrm{V}\big) \end{aligned}$$

$$\begin{aligned} \gamma_c\left(\Theta_1\left(\overset{1000000}{\underset{i=1}{R}} i, j_0|_{j_0=2}\right)\right) &= \overset{1000000}{\underset{i=1}{R}} \gamma_e\left(\Theta_1(i, 2)\right) \\ &= \overset{1000000}{\underset{i=1}{R}}\left(v_i|\mathrm{S}:=\Theta_1(i, 2)|\mathrm{S}\right) \\ &= \left(John|\mathrm{S},\ Judy|\mathrm{S},\ ...,\ Mike|\mathrm{S}\right) \end{aligned}$$

上述对基本BDS的数据对象的检索算子可以扩展到对一般BDS $\Theta^q\left(BDS^q\right)$上的操作。

定义22 在U中对一般BDS $\Theta^q\left(BDS^q\right)$,数据对象的检索算子$\gamma^q_{e/r/c/\omega}\left(\Theta^q\right)$可分为元素检索$\gamma^q\left(\Theta(i, j)\right)$、行检索$\gamma^k_r\left(\Theta\left(i^k_0, \overset{m_k}{\underset{j^k=1}{R}} j^k\right)\right)$、列检索$\gamma^k_c\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}} i^k, j^k_0\right)\right)$和子域检索$\gamma_\omega\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}} i, \overset{m_k}{\underset{j=1}{R}} j\right)\right)$,其可通过迭代导出一系列$q$层自顶向下的基本操作$\overset{1}{\underset{k=q}{R}} \gamma^k_{e/r/c/\omega}\left(\Theta^k_q\right)$,即

$$\gamma^q_{e/r/c/\omega} \triangleq \overset{1}{\underset{k=q}{R}} \gamma^k_{e/r/c/\omega}\left(\Theta^k_q\right)$$

$$\begin{cases} \overset{1}{\underset{k=q}{R}} \gamma^k_e\left(\Theta\left(i^k_0, j^k_0\right)\right) \triangleq \overset{1}{\underset{k=q}{R}}\left[v_k|\mathbb{T}_k:=\Theta^k_q\left(i^k_0|_{i^k_0\in[1,n_k]}, \right.\right. \\ \qquad \left.\left.|_{j^k_0\in[1,m_k]}\right)|\mathbb{T}_{0^k_{j^k_0}}\right] \\ \overset{1}{\underset{k=q}{R}} \gamma^k_r\left(\Theta\left(i^k_0, \overset{m_k}{\underset{j^k=1}{R}} j^k\right)\right) \triangleq \overset{1}{\underset{k=q}{R}}\left[\overset{m_k}{\underset{j^k=1}{R}} \gamma^k_e\left(\Theta\left(i^k_0, j^k\right)\right)\right] \\ \quad = \overset{1}{\underset{k=q}{R}}\overset{m_k}{\underset{j^k=1}{R}}\left[v_k|\mathbb{T}_{j^k}:=\Theta^k_q\left(i^k_0|_{i^k_0\in[1,n_k]}, j^k\right)|\mathbb{T}_{0^k j^k}\right] \\ \overset{1}{\underset{k=q}{R}} \gamma^k_c\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}} i^k, j^k_0\right)\right) \triangleq \overset{1}{\underset{k=q}{R}}\left[\overset{n_k}{\underset{i^k=1}{R}} \gamma^k_e\left(\Theta\left(i^k, j^k_0\right)\right)\right] \\ \quad = \overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}\left[v_{i^k j^k_0}|\mathbb{T}_{j^k_0}:=\Theta^k_q\left(i^k, j^k_0|_{i^k_0\in[1,m_k]}\right)|\mathbb{T}_{0^k_{j^k_0}}\right] \\ \overset{1}{\underset{k=q}{R}} \gamma^k_\omega\left(\left(\Theta\overset{n_k}{\underset{i^k=1}{R}} i^k, \overset{m_k}{\underset{j^k=1}{R}} j^k\right)\right) \triangleq \overset{1}{\underset{k=q}{R}}\left[\overset{n_k}{\underset{i^k=1}{R}}\overset{m_k}{\underset{j^k=1}{R}} \gamma^k_e\left(\Theta\left(i^k, j^k\right)\right)\right] \\ \quad = \overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^k=1}{R}}\overset{m_k}{\underset{j^k=1}{R}}\left[v_{i^k j^k}|\mathbb{T}_{j^k}:=\Theta^k_q\left(i^k, j^k\right)|\mathbb{T}_{0^k j^k}\right] \end{cases} \tag{26}$$

当子域检索的行和列的范围扩展到最大时,

其即转化为对一般BDS的整体检索 $\overset{1}{\underset{k=q}{R}}\gamma_{\Omega}^{k}\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,\ \overset{m}{\underset{j=1}{R}}j\right)\right)$ 或拷贝。

例11　一个如例3所示的一般BDS $\Theta_2^3(BDS_2)$ 的一组特定类型值检索可以根据定义22逐层迭代地用式(26)操作,其实例是例9所示的逆操作。

4.3　BDA数据对象选择算子

对DBS中的数据对象的选择是一个独特的内容匹配过程,其从 $\Theta(BDS)$ 中选择一个或多个符合指定属性的数据项。大数据对象选择算子提供比检索算子更复杂的操作,从而实现一系列基于给定条件表达式而不是特定地址的数据选择。

定义23　在 U 中对基本BDS数据对象的选择算子 $\sigma_{e/r/c/\omega}(\Theta)$,是一个基于给定属性 $\left(p\middle|\mathbb{T}_p\right)$ 匹配而无需指定数据对象的索引 $\left(i\middle|\mathrm{N},j\middle|\mathrm{N}\right)$ 的只读提取操作。其可分为元素选择 $\sigma_e\left(\Theta\left(i_0,\ \overset{m}{\underset{j=1}{R}}j\right)\right)$、行选择 $\sigma_r\left(\Theta\left(i_0,\ \overset{m}{\underset{j=1}{R}}j\right)\right)$、列选择 $\sigma_c\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,j_0\right)\right)$ 和子域选择 $\sigma_\omega\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,\ \overset{m}{\underset{j=1}{R}}j\right)\right)$,即

$$\gamma_{e/r/c/\omega}(\Theta)\triangleq\begin{cases}\sigma_e\left(\Theta(i_0,j_0)\right)\triangleq\Theta\left(i_0\middle|_{i_0\in[1,n]},j_0\middle|_{j_0\in[1,m]}\right)\middle|\mathbb{T}_{0j}\\ \quad\left(\Theta(i_0,j_0)\middle|\mathbb{T}_{0j}\sim p\middle|\mathbb{T}_p\right)\\ \sigma_r\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i,j_0\right)\right)\triangleq\overset{n}{\underset{i=1}{R}}\sigma_e\left(\Theta(i,j_0)\right)\\ \quad=\overset{n}{\underset{i=1}{R}}\left\{\Theta\left(i,j_0\middle|_{j_0\in[1,m]}\right)\middle|\mathbb{T}_{0j_0}\middle|\Theta(i,j_0)\middle|\mathbb{T}_{0j_0}\sim p\middle|\mathbb{T}_p\right\}\\ \sigma_c\left(\Theta\left(i_0,\overset{m}{\underset{j=1}{R}}j\right)\right)\triangleq\overset{m}{\underset{j=1}{R}}\sigma_e\left(\Theta(i_0,j)\right)\\ \quad=\overset{m}{\underset{j=1}{R}}\left\{\Theta\left(i_0\middle|_{i_0\in[1,n]},j\right)\middle|\mathbb{T}_{0j}\middle|\Theta(i_0,j)\middle|\mathbb{T}_{0j}\sim p\middle|\mathbb{T}_p\right\}\\ \sigma_\omega\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}}i,\ \overset{m_k}{\underset{j=1}{R}}j\right)\right)\triangleq\overset{n_k}{\underset{i=1}{R}}\overset{m_k}{\underset{j=1}{R}}\sigma_e\left(\Theta(i,j)\right)\\ \quad=\overset{n_k}{\underset{i=1}{R}}\overset{m_k}{\underset{j=1}{R}}\left\{\Theta(i,j)\middle|\mathbb{T}_{0j}\middle|\Theta(i,j)\middle|\mathbb{T}_{0j}\sim p\middle|\mathbb{T}_p\right\}\end{cases}\tag{27}$$

式中,可选属性是诸如值、类型、模式、尺寸、范围和时间等,由条件运算符 $\sim\triangleq\{=,<,\leqslant,>,\geqslant,\in\}$ 表示。

当子域选择的行和列的范围均为最大时,其即转化为整体选择 $\overset{1}{\underset{k=q}{R}}\sigma_{\Omega}^{k}\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}}i,\ \overset{m_k}{\underset{j=1}{R}}j\right)\right)$。

例12　任给一个基本 $\Theta_1(BDS_1)$ 如例1所示。各种数据对象的基于给定属性 $p\middle|\mathbb{T}_p$ 的选择可以根据定义23操作如下:

$$\sigma_e\left(\Theta_1\left(i_0\middle|_{i_0=2},j_0\middle|_{j_0=1}\right)\right)=\Theta_1(2,1)\middle|\mathrm{N}<\left(p\middle|\mathbb{T}_p=3\middle|\mathrm{N}\right)$$
$$=0000002\middle|\mathrm{N}$$
$$\sigma_r\left(\Theta_1\left(i,j_0\middle|_{j_0=2}\right)\right)=\Theta_1\left(\overset{1000000}{\underset{i=1}{R}}i,2\right)\middle|\mathrm{S}=\left(p\middle|\mathbb{T}_p=J^*\middle|\mathrm{S}\right)$$
$$=\begin{cases}ID\middle|\mathrm{N}=0000001\middle|\mathrm{N},\ UName\middle|\mathrm{S}=John\middle|\mathrm{S},\\ \quad GName\middle|\mathrm{S}=G_{0001}\middle|\mathrm{S},\ Text\middle|\mathrm{TX}=R_t\ 0000001\middle|\mathrm{TX},\\ \quad Voice\middle|\mathrm{A}=R_a\ 0000001\middle|\mathrm{A},\ Photo\middle|\mathrm{F}=R_p\ 0000001\middle|\mathrm{F},\\ \quad Video\middle|\mathrm{V}=R_v\ 0000001\middle|\mathrm{V}\right)\\ ID\middle|\mathrm{N}=0000002\middle|\mathrm{N},\ UName\middle|\mathrm{S}=Judy\middle|\mathrm{S},\\ \quad GName\middle|\mathrm{S}=G_{0301}\middle|\mathrm{S},\ Text\middle|\mathrm{TX}=R_t\ 0000002\middle|\mathrm{TX},\\ \quad Voice\middle|\mathrm{A}=R_a\ 0000002\middle|\mathrm{A},\ Photo\middle|\mathrm{F}=R_p\ 0000002\middle|\mathrm{F},\\ \quad Video\middle|\mathrm{V}=R_v\ 0000002\middle|\mathrm{V}\right)\end{cases}$$
$$\sigma_c\left(\Theta_1\left(\overset{1000000}{\underset{i=1}{R}}i,\ \overset{7}{\underset{j=1}{R}}j\right)\right)=\Theta_1\left(\overset{1000000}{\underset{i=1}{R}}i,\ \overset{7}{\underset{j=1}{R}}j\right)\middle|\mathrm{TX}\geqslant$$
$$\left(p\middle|\mathbb{T}_p=R_t\ 0000002\middle|\mathrm{TX}\right)$$
$$=\overset{1000000}{\underset{i=1}{R}}\Theta_1\left(i,j_0\middle|\mathbb{T}=\mathrm{TX}\right)\middle|\mathrm{TX}\geqslant$$
$$R_t\ 0000002\middle|\mathrm{TX}$$
$$=\left(R_t\ 0000002\middle|\mathrm{TX},\ R_t\ 0000003\middle|\mathrm{TX},\right.$$
$$\left.\cdots,R_t\ 1000000\middle|\mathrm{TX}\right)^{\mathrm{T}}$$

上述对基本BDS的数据对象的选择算子可以扩展到一般BDS $\Theta^q\left(BDS^q\right)$ 上的操作。

定义24　在 U 中对一般BDS $\Theta^q\left(BDS^q\right)$ 的数据对象的选择算子 $\sigma_{e/r/c/\omega}^{q}\left(\Theta^q\right)$,可分为元素选择 $\sigma_e\left(\Theta(i,j)\right)$、行选择 $\sigma_r^k\left(\Theta\left(i_0^k,\ \overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)$、列选择 $\sigma_c^k\left(\Theta\left(\overset{n_k}{\underset{i^k=1}{R}}i^k,j_0^k\right)\right)$ 和子域选择 $\sigma_\omega^k\left(\left(\Theta\overset{n_k}{\underset{i^k=1}{R}}i^k,\ \overset{m_k}{\underset{j^k=1}{R}}j^k\right)\right)$,其可通过迭代导出一系列 q 层自顶向下的基本操作

$\overset{1}{\underset{k=q}{R}}\sigma^{k}_{e/r/c/\omega}\left(\Theta^{k}_{q}\right)$,即

$$\sigma^{q}_{e/r/c/\omega}\left(\Theta^{q}\right)\triangleq\overset{1}{\underset{k=q}{R}}\sigma^{k}_{e/r/c/\omega}\left(\Theta^{k}_{q}\right)$$

$$=\begin{cases}\overset{1}{\underset{k=q}{R}}\sigma^{k}_{e}\left(\Theta\left(\overset{n}{\underset{i=1}{R}}i^{k}_{0},j^{k}_{0}\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\Theta^{k}_{q}\left(i^{k}_{0}\middle|_{i^{k}_{0}\in[1,n_k]},i^{k}_{0}\middle|_{j^{k}_{0}\in[1,m_k]}\right)\middle|\mathbb{T}_{0^{k}j^{k}_{0}}\right.\\ \qquad\left.\left|\Theta^{k}_{q}\left(i^{k}_{0},j^{k}_{0}\right)\right|\mathbb{T}_{0^{k}j^{k}_{0}}\sim p^{k}\middle|\mathbb{T}_{p^{k}}\right]\\ \overset{1}{\underset{k=q}{R}}\sigma^{k}_{r}\left(\Theta\left(\overset{n_k}{\underset{i^{k}=1}{R}}i^{k},j^{k}_{0}\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\left(\overset{n_k}{\underset{i^{k}=1}{R}}\sigma^{k}_{e}\left(\Theta\left(i^{k},j^{k}_{0}\right)\right)\right)\right.\\ =\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^{k}=1}{R}}\left[\Theta^{k}_{q}\left(i^{k},j^{k}_{0}\middle|_{j^{k}_{0}\in[1,m_k]}\right)\middle|\mathbb{T}_{0^{k}j^{k}_{0}}\middle|\Theta^{k}_{q}\left(i^{k},j^{k}_{0}\right)\middle|\mathbb{T}_{0^{k}j^{k}_{0}}\sim p^{k}_{j}\middle|\mathbb{T}_{p^{k}_{j}}\right]\\ \overset{1}{\underset{k=q}{R}}\sigma^{k}_{c}\left(\Theta\left(i^{k}_{0},\overset{m_k}{\underset{j^{k}=1}{R}}j^{k}\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\left(\overset{m_k}{\underset{j^{k}=1}{R}}\sigma^{k}_{e}\left(\Theta\left(i^{k}_{0},j^{k}\right)\right)\right)\right.\\ =\overset{1}{\underset{k=q}{R}}\overset{m_k}{\underset{j^{k}=1}{R}}\left[\Theta^{k}_{q}\left(i^{k}_{0}\middle|_{i^{k}_{0}\in[1,n_k]},j^{k}\right)\middle|\mathbb{T}_{0^{k}j^{k}}\middle|\Theta^{k}_{q}\left(i^{k}_{0},j^{k}\right)\middle|\mathbb{T}_{0^{k}j^{k}}\sim p^{k}_{j}\middle|\mathbb{T}_{p^{k}_{j}}\right]\\ \overset{1}{\underset{k=q}{R}}\sigma^{k}_{\omega}\left(\Theta\left(\overset{n_k}{\underset{i^{k}=1}{R}}i^{k},\overset{m_k}{\underset{j^{k}=1}{R}}j^{k}\right)\right)\triangleq\overset{1}{\underset{k=q}{R}}\left[\left(\overset{n_k}{\underset{i^{k}=1}{R}}\overset{m_k}{\underset{j^{k}=1}{R}}\sigma^{k}_{e}\left(\Theta\left(i^{k},j^{k}\right)\right)\right)\right.\\ =\overset{1}{\underset{k=q}{R}}\overset{n_k}{\underset{i^{k}=1}{R}}\overset{m_k}{\underset{j^{k}=1}{R}}\left[\Theta^{k}_{q}\left(i^{k},j^{k}\right)\middle|\mathbb{T}_{i^{k}j^{k}}\middle|\Theta^{k}_{q}\left(i^{k},j^{k}\right)\middle|\mathbb{T}_{i^{k}j^{k}}\sim p^{k}_{j}\middle|\mathbb{T}_{p^{k}_{j}}\right]\end{cases}\tag{28}$$

当子域选择的行和列的范围均为最大时,其即转化为对一般BDS的整体选择$\overset{1}{\underset{k=q}{R}}\sigma^{k}_{\Omega}\left(\Theta\left(\overset{n_k}{\underset{i=1}{R}}i,\overset{m_k}{\underset{j=1}{R}}j\right)\right)$。

例13 对如例3所示的一般BDS $\Theta^{3}_{2}(BDS_2)$的一组特定选择,可以根据定义24逐层迭代地实现。其实例是例12所示类似单层操作的多层迭代。

4.4 BDA数据定时算子

数据对象的建立或修改时间在BDA中被经常用于数据对象选择、持续时间确定、数值微分和积分。BDS中数据对象的定时等同于对BDS一值域所对应的时域的赋值操作,其中数据对象是与时间相关的带有类型后缀如|T,|D和|TM类型。

定义25 在U中对基本BDS的数据对象的定时算子$\tau(\Theta)$是一种特殊的赋值操作。其对一新创建或更新的数据对象提供一基于当前系统时间$t_c|\mathrm{TM}=t_{sys}|\mathrm{TM}$的时间戳标定,即

$$\tau(\Theta)\triangleq\tau_{e}\left(\Theta\left(i_0,j_0\right)\right)$$

$$=\Theta\left(i_0,j_0\right)\middle|\mathbb{T}_{0j}:=t_c|\mathrm{TM},\ iff\ \mathbb{T}_{0j}=\mathrm{TM}\wedge t_c|\mathrm{TM}=t_{sys}|\mathrm{TM}\tag{29}$$

定义26 在U中对一般BDS $\Theta^{q}\left(BDS^{q}\right)$的数据对象的定时算子$\tau^{q}\left(\Theta^{q}\right)$是$q$层迭代的基本BDS的数据对象时间戳标定,即

$$\tau^{q}\left(\Theta^{q}\right)\triangleq\overset{1}{\underset{k=q}{R}}\tau^{k}_{e}\left(\Theta^{k}_{q}\left(i^{k}_{0},j^{k}_{0}\right)\right)$$

$$=\overset{1}{\underset{k=q}{R}}\left\{\Theta^{k}_{q}\left(i^{k}_{0},j^{k}_{0}\right)\middle|\mathbb{T}_{0j^{k}_{0}}:=t^{k}_{c}|\mathrm{TM}\ \middle|\ t^{k}_{c}|\mathrm{TM}=t_{sys}|\mathrm{TM}\right\}\tag{30}$$

例14 设一系统时间戳$t_c|\mathrm{TM}=t_{sys}|\mathrm{TM}$,所给一般BDS $\Theta^{3}_{2}(BDS_2)$的数据对象的定时可根据定义26操作如下:

$$\tau^{k}\left(\Theta^{3}_{2}(BDS_2)\right)=\tau^{2}_{e}\left(\Theta^{2}_{2}\left(i^{2}_{0}\middle|_{xi^{2}_{0}=1000000},j^{2}_{0}\middle|_{j^{2}_{0}=2}\right)\right)$$

$$=\Theta^{2}_{2}(1000000,2)|\mathrm{TM}:=t_c|\mathrm{TM}$$

4.5 BDA数据微分算子

许多与时间相关的数据对象在BDS中形成时间序列、例如语音信息和视频流。关于同一列时间戳的大数据序列可以通过数值微分来操作。数据微分是一种典型的大数据挖掘方法,用于检测给定时间跨度内的同列数据的动态变化。

定义27 在U中,对基本BDS的数据序列对于时间的微分算子$\delta(\Theta)=\dfrac{\mathrm{d}\Theta(j)}{dj}$是一系列数值对象对其时间序列的数值微分

$$\triangle TStamp|\mathrm{TM}=\Theta(i+1)|\mathrm{SM}.TStamp|\mathrm{TM}-\Theta(i)|\mathrm{SM}.TStamp|\mathrm{TM},\text{即}$$

$$\delta(\Theta)\triangleq\frac{\mathrm{d}\Theta(i,j_0)}{\mathrm{d}i}$$

$$=\overset{n-1}{\underset{i=1}{R}}\frac{\Theta(i+1,j_0)|\mathbb{T}_{0j}-\Theta(i,j_0)|\mathbb{T}_{0j}}{\Theta(i+1,j_0^{'})|\mathrm{TStamp}-\Theta(i,j_0^{'})|\mathrm{TStamp}},\tag{31}$$

$$\mathbb{T}_{j_0^{'}}=\mathrm{TStamp},\quad j_0,j_0^{'}\in[1,m]$$

本操作适用于数值序列类型为$\mathbb{T}_{0j}\in\{\mathbb{N},\mathbb{Z},\mathbb{R},\mathbb{B},\mathbb{H}\mathrm{x}\}$。

定义28 在U中对一般BDS $\Theta^{q}\left(BDS^{q}\right)$的数据序列对于时间的微分算子$\delta^{q}\left(\Theta^{q}\right)=\dfrac{\mathrm{d}\left(\Theta^{q}(j)\right)}{dj}$,是一系列$q$层迭代的单层基本BDS的微分操作,即

$$\delta^q\left(\Theta^q\right)\triangleq\mathop{R}_{k=q}^{1}\delta^k\left(\Theta_q^k\right)=\mathop{R}_{k=q}^{1}\mathop{R}_{i_k=1}^{n_k-1}\frac{\mathrm{d}\left(\Theta_q^k\left(i_k,j_k\right)\right)}{\mathrm{d}i_k}$$

$$=\mathop{R}_{k=q}^{1}\left[\mathop{R}_{i_k=1}^{n_k-1}\frac{\Theta_q^k\left(i_k+1,j_{k0}\right)\left|\mathbb{T}_{0j_k}-\Theta_q^k\left(i_k,j_{k0}\right)\right|\mathbb{T}_{0j_k}}{\Theta_q^k\left(i_k+1,j_{k0}^{'}\right)\left|\mathrm{TStamp}-\Theta_q^k\left(i_k,j_{k0}^{'}\right)\right|\mathrm{TStamp}}\right],$$

$$\mathbb{T}_{0j_k}\in\{\mathbb{N},\mathbb{Z},\mathbb{R},\mathbb{B},\mathbb{H}\},\quad\left|\mathbb{T}_{0j_0}=\mathrm{TStamp},\quad j_{k0},j_{k0}^{'}\in[1,m_k]\right.\tag{32}$$

例 15　对如例3中给出的一般BDS $\Theta_2^3(BDS_2)$,其第2层上的数据时间序列的微分可根据定义28实现如下:

$$\delta^2\left(\Theta_2^2(BDS_2)\right)\triangleq\frac{\mathrm{d}\Theta_2^2\left(i^2,j_0^2\middle|_{j_0^2=3}\right)}{\mathrm{d}i^2}$$

$$=\mathop{R}_{i^2=1}^{1000000-1}\frac{\Theta_2^2\left(i^2+1,3\right)\left|\mathrm{B}-\Theta_2^2\left(i^2,3\right)\right|\mathrm{B}}{\Theta_2^2\left(i^2+1,2\right)\left|\mathrm{TStamp}-\Theta_2^2\left(i^2,2\right)\right|\mathrm{TStamp}}$$

$$=\left(\frac{S_2-S_1}{TM_2-TM_1},\frac{S_3-S_2}{TM_3-TM_2},\cdots,\frac{S_{1000000}-S_{999999}}{TM_{1000000}-TM_{999999}}\right)$$

5　大数据代数的综合算子

基于前文所建立的BDA的数学框架和范例,本节给出一组对大数据系统的综合算子,用以实现对BDS的归纳操作,提取数据的潜在性质和获取数据的应用价值及其可导出的知识。任何对BDS的复杂归纳操作均可表示为这些基本BDA综合算子的代数组合。

定理1　BDS的一般递归超结构性质:大数据系统的超结构Θ^q是一种多维递归层次结构。当终端层Θ^0作为具体数据的类型化元组已给定时,Θ^q中任意第k层可通过下层结构递归地表示,即

$$\Theta^q=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right),\quad\Theta^0=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{j_0=0}^{m_0}d_{i_0j_0}\left|\mathbb{T}_{0j_0}\right.\tag{33}$$

证明　定理1可以根据定义9归纳证明如下:

$$\forall\Theta^0=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{j_0=0}^{m_0}d_{i_0j_0}\left|\mathbb{T}_{0j_0}\right.,$$

$$\Theta^q=\begin{cases}\Theta^1=\Theta^1\left(\Theta^0\right)=\Theta^1\left(\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{j_0=0}^{m_0}d_{i_0j_0}\left|\mathbb{T}_{0j_0}\right.\right)\\ \Theta^2=\Theta^2\left(\Theta^1\left(\Theta^0\right)\right)\\ \cdots\\ \Theta^q=\Theta^q\left(\Theta^{q-1}\left(\ldots\left(\Theta^1\left(\Theta^0\right)\right)\ldots\right)\right)\end{cases}\tag{34}$$

$$=\Theta^q\left(\Theta^{q-1}\left(\ldots\left(\Theta^1\left(\Theta^0\right)\right)\ldots\right)\right)$$

$$=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right)$$

上述定理揭示了作为递归层次结构的BDS不仅可以自顶向下地演绎分析,而且可以从下到上地归纳组合。BDS的这些属性可导出一系列形式化的BDA综合操作。

定义29　BDA的综合算子 $\bullet_s$是在U中对$\Theta(BDS)$的一组归纳操作,用于对大数据系统的挖掘和知识获取,即

$$\bullet_s\triangleq\left\{\iota(\Theta),\lambda(\Theta),\varpi(\Theta),\psi(\Theta)\right\}\tag{35}$$

式中,各算子分别代表归纳、演绎、积分和均衡。

5.1　BDA数据归纳算子

根据定理1,对BDS中数据序列的归纳可以描述为一个自底而上分层的扩展过程。

推论2　BDS的分层抽象原则:在U中对任何$\Theta^q(BDS)$的分层抽象操作$\Theta^q=\mathop{\Uparrow}_{k=0}^{q}\Theta^k$,当最底层的数据序列已给定时,可以通过自底而上的递归归纳导出上层结构,即

$$\Theta^q=\mathop{\Uparrow}_{k=0}^{q}\Theta^k=\mathop{R}_{k=1}^{q}\Theta^k\left(\Theta^{k-1}\right),\ \Theta^0=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{j_0=0}^{m_0}d_{i_0j_0}\left|\mathbb{T}_{0j_0}\right.$$

$$=\Theta^q\left(\Theta^{q-1}\left(\ldots\left(\Theta^1\left(\Theta^0\right)\right)\ldots\right)\right)\tag{36}$$

推论2是系统理论中对系统归纳和信息隐蔽的经典经验原理的严格形式描述[31]。

定义30　大数据归纳原理是一种认知推理过程,用于从U中的$\Theta(BDS)$中的多个递归实例$p(i)$中引出一般模式或规则$\iota(\Theta)$,其令样本预测$\tilde{p}(i)|\mathbb{T}$和平均预测$\overline{p(i)|\mathbb{T}}$之间的误差当样本数足够大时趋于零,即

$$\iota(\Theta)\triangleq\left|\tilde{p}(i)\right|\mathbb{T}-\overline{p(i)|\mathbb{T}}\left|\rightarrow 0\right.$$

$$=\left|\lim_{i\rightarrow k}\left(p\left(s_i|\mathbb{T}\right)|\mathbb{T}\right)-\lim_{i\rightarrow\infty}\left(p\left(s_i|\mathbb{T}\right)|\mathbb{T}\right)\right|\rightarrow 0,\ k\geqslant 3\tag{37}$$

其中归纳操作产生合理的推理外延,将有限的数据序列扩展到更大范围的一般模式。

BDA中的大数据序列归纳算子可以分为3类,称为逻辑、回归和分层归纳操作。

5.1.1 BDS的逻辑归纳

BDS的逻辑归纳是归纳推理模式的逻辑表达形式,即

$$\begin{aligned}&(\exists a\in S,\ a\ \vdash p(a))\wedge\\&(\exists b\in S,\ b\ \vdash p(b))\wedge\\&(\exists r\in S,\ r\ \vdash p(r))\wedge\\&(\exists succ(r)\in S,\ succ(r)\ \vdash p(succ(r)))\\&\Rightarrow\forall x\in S,\ x\ \vdash p(x)\end{aligned}\tag{38}$$

式中,S为一个逻辑变量集;r为在$(3,|\mathrm{S}|)$中任意选择的一个整数数据项。

定义31 在U中对BDS的数据系列的逻辑归纳算子$\iota_l(\Theta)$是一种扩展性的外延推理操作,其从有限序列的实例$p(s_i|s_i\in S)$中外推出一个一般模式$p(S)$,通常$1\leqslant i=(1,2,r,r+1)\leqslant|S|$,其中域$|\mathrm{S}|$的整个范围可外延到$|S|\in[1,\infty)$,即

$$\begin{aligned}&\iota_l(\Theta)\triangleq\underset{i=1,2,r}{\overset{r+1}{R}}(s_i|\mathbb{T}\in S|\mathbb{T}\sqsubset\Theta\ \vdash p(s_i|\mathbb{T})|\mathrm{L}),\\&\quad 2<r<|S|\qquad\qquad\text{// Sample induction}\\&\Uparrow\underset{i=1}{\overset{|S|}{R}}p(s_i|\mathbb{T})|\mathrm{L},\ iff\ \ s_i|\mathbb{T}\in S|\mathbb{T}\sqsubset\Theta\quad\text{// Domain induction}\\&\Uparrow\underset{i=1}{\overset{n}{R}}p(s_i|\mathbb{T})|\mathrm{L},\ iff\ \ s_i|\mathbb{T}\in S|\mathbb{T}\wedge\\&\quad n=|S|\in[3,\infty)\qquad\text{// General induction}\end{aligned}\tag{39}$$

式中,$|\mathrm{L}$表示逻辑变量$\mathrm{L}=\{\mathrm{T},\mathrm{F}\}$的类型后缀。

例16 具有1000个输入的逻辑与门的状态空间可以描述为一个BDS $\Theta_3(BDS_3)=AND\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|\mathrm{L}$,因其所有输入组合所产生的巨大指称状态空间为$|\Theta_3|=|\mathrm{AND}\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|=2^{1000}|\mathrm{b}$。一个在$\mathrm{BDS}_3$上的逻辑归纳操作$\iota_l(\Theta_3)$,可以根据定义31导出该与门的一般功能如下:

$$\begin{aligned}&\iota_l(\Theta_3)=\iota_l\left(AND\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|\mathrm{L}\right)\\&\Uparrow AND\left(\underset{i=1}{\overset{n}{R}}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{n}d_i|\mathrm{b},\ \forall n\in(1,\infty)\\&\Uparrow AND\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b}\\&\Uparrow\begin{cases}AND\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|\mathrm{L}=1,\ \forall\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}=1\\AND\left(\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}\right)|\mathrm{L}=0,\ \exists\underset{i=1}{\overset{1000}{R}}d_i|\mathrm{b}=0\end{cases}\end{aligned}$$

式中,$|b$表示二进数位的类型后缀。

5.1.2 BDS的回归归纳

BDS的回归归纳是归纳推理模式的代数表达形式。其基于多项式回归对大数据所隐含的一般多项式函数做数值拟合。该操作的数据对象适用于实数或与其兼容的数据类型。通过对给定样本数据序列的实例$\underset{i=1}{\overset{n}{R}}(x_i|\mathbb{R},y_i|\mathbb{R})$进行有限代数操作,回归归纳得以实现。

定义32 在U中对BDS的数据系列的回归归纳算子$\iota_r(\Theta)$通过一m阶多项式,$y=\sum_{i=0}^{m}a_ix^i=a_mx^m+\ldots+a_3x^3+a_2x^2+a_1x+a_0$的最小方差回归对$\Theta(\mathrm{BDS})$中的一数据序列进行拟合。通常$m\leqslant4$,$n=m+1$,最小方差$\mathrm{SLE}^4(\Theta_4)$为

$$E_{ls}^4=\underset{i=0}{\overset{4}{R}}\frac{\partial}{\partial a_i}\sum_{i=1}^{5}\left[y_i-\left(a_4x_i^4+a_3x_i^3+a_2x_i^2+a_1x_i+a_0\right)\right]^2=0,$$

即

$$\begin{aligned}&\mathrm{SLE}^4\triangleq\mathbf{SA}=\mathbf{M}\\&\triangleq\begin{bmatrix}n&S_x&S_{x^2}&S_{x^3}&S_{x^4}\\S_x&S_{x^2}&S_{x^3}&S_{x^4}&S_{x^5}\\S_{x^2}&S_{x^3}&S_{x^4}&S_{x^5}&S_{x^6}\\S_{x^3}&S_{x^4}&S_{x^5}&S_{x^6}&S_{x^7}\\S_{x^4}&S_{x^5}&S_{x^6}&S_{x^7}&S_{x^8}\end{bmatrix}\begin{bmatrix}a_0\\a_1\\a_2\\a_3\\a_4\end{bmatrix}=\begin{bmatrix}S_y\\S_{xy}\\S_{x^2y}\\S_{x^3y}\\S_{x^4y}\end{bmatrix}\end{aligned}\tag{40}$$

其中 $S_{x^k}=\sum_{i=1}^{n}x_i^k$, $S_{y^k}=\sum_{i=1}^{n}y_i^k$,

$$S_{x^ky^{k'}}=\sum_{i=1}^{n}x_i^k\,y_i^{k'},\ 1\leqslant k\leqslant2m$$

因此,$\mathbf{A}=\mathbf{S}\backslash\mathbf{M}=[a_0\ \ a_1\ \ a_2\ \ a_3\ \ a_4]^{\mathrm{T}}$

即:$\iota_r(\Theta)\triangleq y=a_4x^4+a_3x^3+a_2x^2+a_1x+a_0$

例17 令一社交网络中的意见分布大数据系统$\Theta_4(BDS_4)$如式(41)所示。当给定m=2,对它的多项式回归归纳操作$\mathrm{SLE}^2(\Theta_1)$,可以根据定义32导出如下:

$$\Theta_4(BDS_4)=\underset{i=0}{\overset{5}{R}}\ \underset{j=0}{\overset{3}{R}}d_{ij}|\mathbb{T}_{0j}$$

$$=\left[\begin{array}{c|c|c|c} \theta_0(BDS_4) & Theme|\mathrm{S} & Spectrum|\mathrm{R} & Supporters|\mathrm{N} \\ \hline \kappa_1 & T_1 & 10 & 5000 \\ \kappa_2 & T_2 & 30 & 30000 \\ \kappa_3 & T_3 & 50 & 10000 \\ \kappa_4 & T_4 & 70 & 20000 \\ \kappa_5 & T_5 & 90 & 6000 \end{array}\right] \quad (41)$$

$$\mathrm{SLE}^2(\Theta_4)\triangleq \mathbf{SA}=\mathbf{M},\ T_i\in[10,\ 50]$$

$$=\begin{bmatrix} n & S_x & S_{x^2} \\ S_x & S_{x^2} & S_{x^3} \\ S_{x^2} & S_{x^3} & S_{x^4} \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}=\begin{bmatrix} S_y \\ S_{xy} \\ S_{x^2y} \end{bmatrix}$$

$$=\begin{bmatrix} 3 & 90 & 3500 \\ 90 & 3500 & 153000 \\ 3500 & 153000 & 7070000 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}=\begin{bmatrix} 45000 \\ 1450000 \\ 52500000 \end{bmatrix}$$

得到 $\mathbf{A}=\mathbf{S}\backslash\mathbf{M}=[a_0\ \ a_1\ \ a_2]^{\mathrm{T}}$

$$=[-24375\ \ 3500\ \ -56]^{\mathrm{T}}$$

以上四归归纳得到

$$\iota_r(\Theta_4)\triangleq y=-56x^2+3500x-24375$$

对*BDS*的回归归纳操作构成现代人工智能即神经网络和深度学习的数学基础。

推论3　基于*BDS*的神经网络的性质:在U中对任何*BDS*的回归归纳操作取决于数据样本的特征蕴含,并非它的量或尺度。

根据推论3,用于训练机器学习的样本不是越大越好,而在于样本数据的特征代表性和独立性。

传统机器学习理论及方法极大地依赖于数据所蕴含的特定样本空间。对于神经网络的这一基本性质的研究导出了下述原理。

推论4　基于回归归纳的机器学习结果的数学本质:在U中对任何*BDS*的回归归纳学习是对一有限集的特定对象在给定范畴的特殊解,而非一般解。

推论4所述回归归纳的数学性质揭示了传统机器学习方法的一个先天限制条件,即不同的训练样本决定不同问题空间上的机器学习结果。因此基于对面部数据的训练不能取代对其他种类图像的训练,例如指纹或花朵。基于特定数据样本的训练,无法推广到它的超集,例如从特定人群面部数据的训练结果到猴脸识别。

5.1.3　BDS的分层归纳

在定义31或定义32中的大数据归纳算子可被应用于一般复杂BDS的每个特定层次,从而导出BDS的分层归纳算子。

定义33　在*U*中对一般BDS $\Theta^q(BDS^q)$的数据系列的分层归纳算子$\iota(\Theta^q)$是一递归的自底向上逐层操作的基本数据归纳运算,以便在Θ^q中提取分层模式的一般外推,即

$$\begin{aligned}\iota(\Theta^q)&=\mathop{R}_{k=1}^{q}\iota^k(\Theta_q^k),\ \exists\Theta_q^0\\&=\Uparrow_{k=1}^{q}\iota^k\left(\Theta_q^k\left(\Theta_q^{k-1}\right)\right)\\&=\iota^q\left(\iota^{q-1}\left(\ldots\left(\iota^1(\Theta_q^0)\right)\ldots\right)\right)\end{aligned}\quad (42)$$

例18　重用例16中的大数据系统$\Theta_3(BDS_3)$,在一般BDS中从数据集到知识和智能模式的分层归纳可以根据定义33实现如下:

$$\iota(\Theta_3)=\mathop{R}_{k=1}^{3}\iota_l^k(\Theta_3^k),\ \Theta_3^0(BDS_3)=\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}$$

$$\Uparrow\mathrm{Data}(\mathbb{D}):\quad \iota_l^0(\Theta_3^0)=2^{\left|\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right|}2^{1000}|\mathrm{b}$$

$$\Uparrow\mathrm{Information}(\mathbb{I}):\quad \iota_l^1(\Theta_3^1)=\left|\left|\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right|\right|=1000|\mathrm{b}$$

$$\Uparrow\mathrm{Knowledge}(\mathbb{K}):\quad \iota_l^2(\Theta_3^0)=$$

$$\begin{cases} AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=1,\ \forall\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=1 \\ AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=0,\ \exists\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=0 \end{cases}$$

$$\Uparrow\mathrm{Intelligence}\left(\dot{\mathbb{I}}\right):\iota_l^3(\Theta_3^3)=AND\left(\mathop{R}_{i=1}^{n}d_i|\mathrm{b}\right)|\mathrm{L}$$

$$=\bigwedge_{i=1}^{n}d_i|\mathrm{b},\ \forall n\in(1,\infty)$$

上述推论2~推论4和定义30~定义33中的实例研究揭示了人类智能在大数据挖掘中对学习和知识生成的不可或缺的归纳能力[1,3,7-8,11-13,16,40,45-47]。人类智能是一种知识发展和创新的极其有效和快速收敛的归纳机制,其中数据仅仅是现实世界的几乎无限状态空间中的事实材料和有限实例。归纳认知过程是人类智能最典型和最强大的推理能力,它将大数据转化为机器几乎在可见的将来无法实现的知识和智慧[20,48-50]。

5.2　BDA数据演绎算子

BDA中的数据序列的演绎是数据序列归纳的

逆操作。根据定理1,BDA中数据序列的演绎可以描述为一个自顶向下的分层细化过程。

推论5 BDS的分层细化原则;在U中对任何$\Theta^q(BDS)$的分层细化操作$\Theta^q=\Downarrow_{k=q}^{0}\Theta^k$,可以通过自顶向下的递归细节演绎导出下层结构,即

$$\Theta^q=\Downarrow_{k=q}^{0}\Theta^k=\mathop{R}_{k=q}^{1}\Theta^k\left(\Theta^{k-1}\right),\ \Theta^0=\mathop{R}_{i_0=0}^{n_0}\mathop{R}_{j_0=0}^{m_0}d_{i_0j_0}\left|\mathbb{T}_{0j_0}\right.$$
$$=\Theta^q\left(\Theta^{q-1}\left(...\left(\Theta^1\left(\Theta^0\right)\right)...\right)\right)\quad(43)$$

该推论是系统科学理论中系统演绎和自顶向下设计的经验原则的严格形式表述[31]。

BDA中的大数据序列演绎算子可以分为3类,称为逻辑、插值和层次演绎操作。

5.2.1 BDS的逻辑演绎

定义34 在U中对BDS的数据序列的逻辑演绎算子$\lambda_l(\Theta)$是一种数据细化操作,其从一个$\Theta(BDS)$的一般模式$p(S)$导出一个蕴含其中的具体范例$p(s_i|s_i\in S)$,即

$$\lambda_l(\Theta)\triangleq\mathop{R}_{i=1}^{|S|}\left(s_i|\mathbb{T}\in S|\mathbb{T}\sqsubset\Theta\ \vdash p\left(s_i|\mathbb{T}\right)|\mathrm{L}\right)$$
$$\Downarrow a|\mathbb{T}\ \vdash p\left(a|\mathbb{T}\right)|\mathrm{L},\ \forall a|\mathbb{T}\in S|\mathbb{T}\quad(44)$$

其所对应的演绎模式的逻辑表达如下所示:

$$\forall s\in S,\ s\ \vdash p(s)\Rightarrow\exists a,\ a\ \vdash p(a),\ iff\ a\in S\quad(45)$$

例19 令$\Theta_3(BDS_3)=AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}$,见例16。所给与门的大数据序列的逻辑演绎可以根据定义34从其高层的一般模式推出下层的具体功能行为如下:

$$\lambda_l(\Theta_3)=\lambda_l\left(AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}\right)$$
$$\Downarrow AND\left(\mathop{R}_{i=1}^{n}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{n}d_i|\mathrm{b},\ \forall n\in N=(1,\infty)$$
// General deduction
$$\Downarrow AND\left(\mathop{R}_{i=1}^{k}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{k}d_i|\mathrm{b},\ k\in N$$
// Domain deduction
$$\Downarrow AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b},\ k=1000\in N$$
// Refined deduction
$$=\begin{cases}AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b}=1,\ \forall\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=1\\ AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b}=0,\ \exists\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=0\end{cases}$$

5.2.2 BDS的插值演绎

BDS上的插值演绎操作是基于一已知如定义32所示的大数据序列的数值多项式。插值演绎用于创建一优化的数据序列特征函数,其中插值$\left(x_{int}^{'},y_{int}^{'}\right),1<int<n$,是在数据序列$\mathop{R}_{i=1}^{n}(x_i,y_i)$上导出的一对或一组特定值对偶。

定义35 在U中对BDS的数据系列的插值演绎算子$\lambda_i(\Theta)$,通过一个在指定域$x\in[x_{\min},x_{\max}]$的m阶多项式$p(x)=\sum_{i=0}^{m}a_ix^i$中的插值$(x_\psi,y_\psi)$导出一优化的数据序列特征函数,即

$$\lambda_i\left(\Theta\left(p(x)\right)\right)\triangleq p\left(x|\mathrm{R}\right)|\mathrm{R}$$
$$=\sum_{i=0}^{m}a_ix^i,\ x^i\in X|\mathrm{R},\ a_i\in\mathop{R}_{i=0}^{n+1}a_i|\mathrm{R}$$
$$\Downarrow y_\psi|\mathrm{R}=p_\psi\left(x_\psi|\mathrm{R}\right)|\mathrm{R}=\sum_{i=0}^{m}a_ix_\psi^i,\ x_\psi\in X|\mathrm{R}\quad(46)$$

例20 根据定义35,对式(41)所给的数据序列$\Theta_4(BDS_4)$进行插值演绎操作。假设m=2,该插值演绎推导结果为

$$\lambda_i\left(\Theta_4\left(BDS_4\right)\right)=\mathop{R}_{x=10}^{50}\left(y=-56x^2+3500x-24375\right),$$
$$x|\mathrm{R}\in\Theta_4|\mathrm{R}=[10,50]$$
$$\Downarrow\forall x_\psi|\mathrm{R}=20\in\Theta_4|\mathrm{R},$$
$$y_\psi=-56x_\psi^2+3500x_\psi-24375$$
$$=-56\times20^2+3500\times20-24375$$
$$=23225$$

5.2.3 BDS的层次演绎

在定义34或定义35中的大数据演绎算子可被应用在一般复杂BDS的每个特定层次,从而导出BDS的分层演绎算子。

定义36 在U中对BDS $\Theta^q\left(BDS^q\right)$的数据系列的层次演绎算子$\lambda_q\left(\Theta^q\right)$是一递归的自顶向下逐

层操作的基本BDS数据演绎运算,以便在 Θ^q 中获取由高层一般模式导出的低层特例,即

$$\begin{aligned}\lambda\left(\Theta^q\right)&=\mathop{R}_{k=q}^{1}\lambda^k\left(\Theta_q^k\right),\ \Theta_q^0\ \text{已知}\\&=\Downarrow_{k=q}^{1}\lambda^k\left(\Theta_q^k\left(\Theta_q^{k-1}\right)\right)\\&=\lambda^q\left(\lambda^{q-1}\left(...\left(\lambda^1\left(\Theta_q^1\left(\Theta_q^0\right)\right)\right)...\right)\right)\end{aligned}\tag{47}$$

例21　基于例16中的大数据系统 $\Theta_3(BDS_3)$ 在一般BDS中从智能模式,知识到数据集的分层演绎可以根据定义36操作如下:

$$\lambda\left(\Theta_3^3\right)=\mathop{R}_{k=3}^{1}\lambda_l^k\left(\Theta_3^k\right),\ \Theta_3^0(BDS_3)=\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}$$

$$\Downarrow\lambda_l^3\left(\Theta_3^3\right)=AND\left(\mathop{R}_{i=1}^{n}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{n}d_i|\mathrm{b},\ \forall n\in(1,\infty)$$

$$\Downarrow\lambda_l^2\left(\Theta_3^2\right)=AND\left(\mathop{R}_{i=1}^{k}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{k}d_i|\mathrm{b},\ \forall k<n$$

$$\begin{aligned}\Downarrow\lambda_l^1\left(\Theta_3^1\right)&=\Theta_3^1\left(\Theta_3^0\right)=AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}\\&=\bigwedge_{i=1}^{1000}d_i|\mathrm{b},\ \forall k<1000\end{aligned}$$

$$=\begin{cases}AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b}=1,\ \forall\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=1\\AND\left(\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}\right)|\mathrm{L}=\bigwedge_{i=1}^{1000}d_i|\mathrm{b}=0,\ \exists\mathop{R}_{i=1}^{1000}d_i|\mathrm{b}=0\end{cases}$$

5.3　BDA数据积分算子

大数据积分是对BDS的综合操作,以便导出给定数据序列的综合加权和。BDS上的其他综合操作,例如数据的统计平均值,数据对整体的比率,以及数据系统的均衡,均依赖于大数据序列的积分操作。

定义37　在U中对 $\Theta(BDS)$ 的积分算子 $\omega(\Theta)$ 是一组符合给定条件 $p|\mathbb{T}$ 的数据序列的数值加权和,可分为行积分 $\int_1^m\Theta_r(i_0,j)\mathrm{d}j$ 、列积分 $\int_1^n\Theta_c(i,j_0)\mathrm{d}i$、子域积分 $\int_{n_1}^{n_2}\int_{m_1}^{m_2}\Theta(i_s,j_s)\mathrm{d}j_s\mathrm{d}i_s$ 和全域积分 $\int_1^n\int_1^m\Theta(i,j)\mathrm{d}j\mathrm{d}i$,即(48)。式中,条件关系符 $\sim\triangleq(=,\leqslant,\geqslant,\in)$ 用以确定在通用数值数据类 $\mathbb{T}\in\{\mathbb{N},\mathbb{Z},\mathbb{R},\mathbb{B},\mathbb{H}\}$ 之间的等价性。而其他特定于问题的关系运算符则依赖于给定 $d_{ij}|\mathbb{T}$ 的类型。

$$\varpi(\Theta)\triangleq\begin{cases}\int_1^m\Theta_r(i_0,j)\mathrm{d}j\triangleq\sum_{j=1}^{m}\left\{d_{ij}|\mathbb{T}_j\cdot\Delta d_j\middle|d_{ij}|\mathbb{T}_j\sim p|\mathbb{T}\right\}\\\int_1^n\Theta_c(i,j_0)\mathrm{d}i\triangleq\sum_{i=1}^{n}\left\{d_{ij}|\mathbb{T}_j\cdot\Delta d_j\middle|d_{ij}|\mathbb{T}_j\sim p|\mathbb{T}\right\}\\\int_{n_1}^{n_2}\int_{m_1}^{m_2}\Theta(i_s,j_s)\mathrm{d}j_s\mathrm{d}i_s\triangleq\sum_{n_1=1}^{n_2}\left(\sum_{m_1=1}^{m_2}\left(d_{i_sj_s}|\mathbb{T}_{0j_s}\cdot\Delta d_{j_s}\right)\right)\cdot\Delta\mathrm{d}i_s\\\int_1^n\int_1^m\Theta(i_s,j)\mathrm{d}j\mathrm{d}i\triangleq\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left(d_{ij}|\mathbb{T}_j\cdot\Delta dj\right)\right)\cdot\Delta d_i\end{cases}\tag{48}$$

例22　对如例3所给一般 BDS $\Theta_2^3(BDS_2)$ 的第2层第3域数据序列的数值积分,可以根据定义37实现如下:

$$\begin{aligned}\varpi\left(\Theta_2^3(BDS_2)\right)&=\int_1^n\Theta_3^2\left(i_2,j_0^2\middle|_{j_0^2=3}\right)\mathrm{d}i^2,\ n=1000\\&=\sum_{i^2=1}^{1000}\left(d_{i^23}|\mathrm{B}\cdot\Delta d_{i^2}\right),\ \Delta d_{i^2}|\mathrm{N}\equiv1.0\\&=\sum_{i^2=1}^{1000}d_{i^23}|\mathrm{B}\end{aligned}$$

定义38　在U中对 $\Theta(BDS)$ 的数据序列在时间上的积分算子 $\omega_t(\Theta_t)=\int_{t_a}^{t_b}\Theta\left(i,j_0\middle|_{j_0=t}\right)\mathrm{d}t$,是一列时间序列中数据对象在时间 $j_0|\mathrm{TM}=TStamp|\mathrm{TM}\in[t_a|\mathrm{TM},t_b|\mathrm{TM}]$ 上的数值积分,即

$$\omega_t(\Theta_t)\triangleq\begin{cases}\int_{t_a}^{t_b}\Theta_c\left(i,j_0|\mathrm{TM}\right)\mathrm{d}t\triangleq\sum_{t=t_a}^{t_b}\left(d_{ij_0}|\mathbb{T}_{j_0}\cdot\Delta t_{j_0'}|\mathbb{T}_{j_0'}\right),\\\quad\mathbb{T}_{j_0'}=\mathrm{TStamp}|\mathrm{TM}\in[t_a|\mathrm{TM},t_b|\mathrm{TM}]\\\int_{t_0}^{t_b}\Theta_c\left(i,j_0|\mathrm{TM}\right)\mathrm{d}t\triangleq\sum_{t=t_0}^{t_b}\left(d_{ij_0}|\mathbb{T}_{j_0}\cdot\Delta t_{j_0'}|\mathbb{T}_{j_0'}\right),\\\quad\mathbb{T}_{j_0}=\mathrm{TStamp}|\mathrm{TM}\leqslant t_b|\mathrm{TM},\\\quad t_0|\mathrm{TM}=SysInitialTime|\mathrm{TM}\\\int_{t_a}^{t_c}\Theta_c\left(i,j_0|\mathrm{TM}\right)\mathrm{d}t\triangleq\sum_{t=t_a}^{t_c}\left(d_{ij_0}|\mathbb{T}_{j_0}\cdot\Delta t_{j_0'}|\mathbb{T}_{j_0'}\right),\\\quad\mathbb{T}_{j_0}=\mathrm{TStamp}|\mathrm{TM}\geqslant t_a|\mathrm{TM},\\\quad t_c|\mathrm{TM}=CurrentSysTime|\mathrm{TM}\end{cases}\tag{49}$$

其中,积分区间表示一特定的时间段 $[t_a,t_b]$。其可是从系统初始化开始到当前的时区 $[t_0,t_b]$,或从给定点到当前时间的时区 $[t_a,t_c]$。

例23 对如例3所给一般BDS $\Theta_2^3(BDS_2)$ 的第2层第3域数据序列对于时间TStamp|TM的数值积分，可以视为单层数值序列积分。根据定义38导出结果如下：

$$\begin{aligned}\omega_t\left(\Theta_2^3\left(BDS_2^3\right)\right)&=\int_{t_a}^{t_b}\Theta_2^2\left(i^2,j_0^2|\mathrm{TM}\right)\mathrm{d}t\\&\quad(\text{其中}t_a=5,\ t_b=126,\ j_0^2|\mathrm{TM}=3)\\&=\sum_{t=5}^{126}\left(d_{i3}|\mathbb{T}_3\cdot\Delta t_3\right),\ \Delta t_3\equiv 1.0\\&=\sum_{i^2=5}^{126}d_{i^23}|\mathrm{B}\end{aligned}$$

基本数据序列在 $\Theta(BDS)$ 上的数值积分可以逐层扩展到一般 $\Theta^q\left(BDS^q\right)$ 上的数值积分，其定义如下。

定义39 在U中对于一般 $\Theta^q\left(BDS^q\right)$ 的数据序列在时间上的数值积分算子 $\mathop{R}_{k=1}^{q}\omega_t^k\left(\Theta_t^k\right)$，是一系列递归的单层基本时间序列数值积分，即

$$\begin{aligned}\mathop{R}_{k=1}^{q}\varpi_t^k\left(\Theta^k\right)&\triangleq\mathop{R}_{k=1}^{q}\left[\int_1^{n_k}\int_1^{m_k}\Theta^k\left(i_k,j_k\right)\mathrm{d}j_k\mathrm{d}i_k\right]\\&=\mathop{R}_{k=1}^{q}\left[\sum_{i=1}^{n_k}\left(\sum_{j=1}^{m_k}\left(d_{i_kj_k}|\mathbb{T}_{j_k}\cdot\Delta d_{j_k}\right)\right)\cdot\Delta d_{i_k}\right]\end{aligned}\tag{50}$$

通常，除非每个目标数据序列或某一层的行的类型适合于数值积分，否则不必在一般BDS中的所有层上进行数值积分。

5.4 BDA数据均衡算子

在以人为中心的BDS中，诸如社会网络、知识库和投票系统，综合观点提取的基本机制是发现所有意见特征分布的均衡点。因此，隐含于BDS中综合观点的代表性意见不是简单的最大值，也不是常规认为的加权平均值。相反，它是代表意见均衡的数据分布的特征点ξ，见如下定义[22]。

定义40 给定U中的一个 $\Theta(BDS)$ 集体观点的均衡点 $\psi(\Theta)$ 是分布在意见频谱 $x\in[1,n]$ 上的一个特征点ξ，$1<\xi<n$，其使加权意见 $p(x)$ 在点ξ的两侧达到平衡，即

$$\psi(\Theta)\triangleq\xi,\ iff\ \int_1^{\xi^-}p(x)\mathrm{d}x=\int_{\xi^+}^{n}p(x)\mathrm{d}x\tag{51}$$

其中 $\int_1^{\xi^-}p(x)\mathrm{d}x+\int_{\xi^+}^{n}p(x)\mathrm{d}x=\int_1^{n}p(x)\mathrm{d}x$.

式(51)中给出的均衡观点 $\psi(\Theta)$ 可以使用数值积分迭代方法来求解。一种实用的大数据特征观点均衡检测算法可见[21]。

定义41 在U中对 $\Theta(BDS)$ 的均衡检测算子 $\psi(\Theta)$ 是点ξ在两边权重之和的迭代匹配在[1, m]达到均衡时的值：

$$\begin{aligned}&\psi\left(\Theta\left(\mathop{R}_{\xi=1}^{n}i,j_0\right)\right)\triangleq\xi+0.5,\quad j_0\in[1,m],\\&iff\ \mathop{R}_{\xi=1}^{n}\left\{\sum_{i=1}^{\xi}\Theta(i,j_0)\approx\sum_{i=\xi+1}^{n}\Theta(i,j_0)\right\}\end{aligned}\tag{52}$$

其中，$\Theta(i,j_0)$ 中的列j_0是根据给定意见分布谱从左到右在[1，m]中排序，或反之。

在定义41中，BDS的某列之间的数值解取 $\psi(\Theta)=\xi+0.5$，因为离散意见分布谱的索引是以单位步长递增的。值得注意的是，BDA中的意见均衡操作依赖于数据属性，并非所有大数据系统中的数据列都适用此操作。

例24 如式(41)所给大数据系统，对 $\Theta(BDS_4)$ 中集体意见在分布谱[10, 90]内的一组分布数据，意见均衡 $\psi(BDS_4)$ 可以根据定义41确定如下：

$$\begin{aligned}\psi(BDS_4)&=\psi_\Theta^3\left(\Theta\left(\mathop{R}_{\xi=1}^{5}i,j_0\big|_{j_0=3}\right)\right)\\&=\xi+0.5\approx 40\\\text{其中}\sum_{i=10}^{30}\Theta\left(i,j_0\big|_{j_0=3}\right)&=35000\approx\\\sum_{i=50}^{90}\Theta\left(i,j_0\big|_{j_0=3}\right)&=36000\end{aligned}\tag{53}$$

上述结果表明，在所给社交网络中分布意见的整体特征均衡点处于意见谱的左中部。进一步增加的大数据将可提高大规模社交网络的BDS中的均衡估计的准确性。定义41所给均衡算子可以扩展到一般BDS的递归层次均衡检测。BDS特征观点的质心揭示了大数据社交网络分析中的一个本质属性。在一系列历史大数据的基础上，可以建立意见平衡基准，用以严格分析社会网络中大规模集体意见中动态均衡质心的变化趋势。因此，大数据特征均衡是大型社交网络动态特征挖掘的一种重要方法。

6　结论

描述了一项关于大数据科学的基础研究,称为大数据代数(BDA),它为大数据工程提供了一种严格而有效的大数据分析方法和大数据系统软件的基础算法。所建立的形式大数据系统(BDS)和代数算子集的数学模型构成了一种BDA的指称数学结构。BDA中的每个算子都代表了一个代数模型,以便实现严格的BDS操作算法。BDA提供了一种有效的数学手段和方法论,用于形式地处理科学、工程和社会中普遍的、前所未有的BDS规模和复杂性的挑战。BDA所揭示的大数据原理和形式模型不仅解释了BDS性质的基本数学结构,而且提供了在大数据分析中严格操纵BDS的有效算法。研究发现大数据系统是一个递归类形化超结构(RTHS)。基于RTHS所创建的新型数学运算使BDA能够应用于大数据工程、机器学习和知识提取的广泛领域。

致谢　此项工作获得加拿大自然科学和工程研究委员会(NSERC)发现基金和清华大学大数据系统软件国家工程重点实验室的支持。陆建华院士和孙家广院士对本文撰写提供了精心指导,王建民和刘琳对专题文章的组织给予了帮助。

参考文献(References)

[1] Chicurel M. Databasing the brain[J]. Nature, 2000, 406 (6798): 822-825.

[2] Codd E F. A relational model of data for large shared data banks[J]. Communications of the ACM, 1970, 13(1): 377-387.

[3] Debenham J K. Knowledge systems design[M]. New York: Prentice Hall, 1989.

[4] Hassanien A E , Azar A T, Snasel V, et al. Big data in complex systems: Challenges and opportunities[M]. Berlin: Springer, 2015.

[5] Jacobs A. The pathologies of big data[J]. Queue, 2009, 7 (6): 10.

[6] Mashey J R. Big data and the next wave of infrastress[J]. SGI, 1998: 1-46.

[7] McCarthy J, Minsky M L, Rochester N, et al. Proposal for the 1956 dartmouth summer research project on artificial intelligence[R/OL]. [2019-10-31]. http://www.formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

[8] McCulloch W S. Embodiments of mind[M]. Cambridge: MIT Press, 1965.

[9] McKinsey B, Gartner D. Big Data means high value, not just volume[J]. Computer Weekly, 2011, 6: 1-2.

[10] Snijders C, Matzat U, Reips U D. ´Big data´: Big gaps of knowledge in the field of internet[J]. International Journal of Internet Science, 2012, 7: 1-5.

[11] Shannon C E . A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27: 379-423, 623-656.

[12] Tucker A B. The computer science and engineering handbook[J]. New York: CRC Press, 1992.

[13] Turing A M. Computing machinery and intelligence[J]. Mind, 1950, 59: 433-460.

[14] Ullman J D, Widom J. A first course in database systems [M]. New York: Prentice Hall, Inc., 1997.

[15] von Neumann J. The computer and the brain[M]. New Haven: Yale University Press, 1958.

[16] Wang Y. On cognitive informatics[J]. Brain and Mind, 2003, 4(3): 151-167.

[17] Wang Y. In search of denotational mathematics: Novel mathematical means for contemporary intelligence, brain, and knowledge sciences[J]. Journal of Advanced Mathematics and Applications, 2012, 1(1): 4-25.

[18] Wang Y. Software science: On general mathematical models and formal properties of software[J], Journal of Advanced Mathematics and Applications, 2014, 3(2): 130-147.

[19] Wang Y. Keynote: Big data algebra: A rigorous approach to big data analytics and engineering[C]//17th International Conference on Mathematical and Computational Methods in Science and Engineering (MACMESE ´15), Kuala Lumpur, 2015: 2.

[20] Wang Y, Tunstel E. Emergence of abstract sciences and transdisciplinary advances in systems, man, and cybernetics[J]. IEEE System, Man and Cybernetics Magazine, 2019, 5(2): 12-19.

[21] Wang Y, Wiebe V J. Big data analytics on the characteristic equilibrium of collective opinions in social networks [J]. International Journal of Cognitive Informatics and Natural Intelligence, 2014, 8(3): 27-42.

[22] Wang Y. Formal cognitive models of data, information, knowledge, and intelligence[J]. WSEAS Transactions on

Computers, 2015, 14: 770-781.

[23] Zadeh L A. Fuzzy sets[J]. Information & Control, 1965, 8 (3): 338-353.

[24] Zadeh L A. Fuzzy logic and approximate reasoning[J]. Synthese, 1975, 30(3-4): 407-428.

[25] Wang Y. On cognitive foundations of big data science and engineering[C]//Proceedings of 15th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC´16). Stanford: IEEE CS Press, 2016: 252-259.

[26] Wang Y. On the informatics laws and deductive semantics of software[J]. IEEE Transactions on Systems, Man, and Cybernetics (Part C), 2006, 36(2): 161-171.

[27] Hartmanis J. On computational complexity and the nature of computer science, 1994 turing award lecture[J]. Communications of the ACM, 1994, 37(10): 37-43.

[28] Wang Y. On Abstract Intelligence: Toward a unified theory of natural, artificial, machinable, and computational intelligence[J]. International Journal of Software Science and Computational Intelligence, 2009, 1(1): 1-17.

[29] Wang Y.The theory of fuzzy arithmetic in the extended domain of fuzzy numbers[J]. Journal of Advanced Mathematics and Applications, 2014, 3(2): 165-175.

[30] Wang Y. On mathematical theories and cognitive foundations of information[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2015, 9(3): 41-63.

[31] Wang Y. Towards the abstract system theory of system science for cognitive and intelligent systems[J]. Journal of Complex and Intelligent Systems, 2015, 1(1): 1-22.

[32] Wang Y. On probability algebra: classic probability theory revisited[J]. WSEAS Transaction on Mathematics, 2016, 15: 550-565.

[33] Cardelli L, Wegner P. On understanding types, data abstraction, and polymorphism[J]. ACM Computing Surveys, 1985, 17(4): 471-522.

[34] Chapra S C, Canale R P. Numerical methods for engineers with software and programming applications[M]. Boston: McGraw-Hill, 2002.

[35] Gowers T. The princeton companion to mathematics[M]. Princeton: Princeton University Press, 2008.

[36] Guttag J V.Abstract data types and the development of data structures[J]. Communications of the ACM, 1977, 20 (6): 396-404.

[37] Lewis H R, Papadimitriou C H, Elements of the theory of computation[M]. New York: Prentice Hall, 1998.

[38] Mitchell J C. Type systems for programming languages [M]//van Leeuwen J. Handbook of Theoretical Computer Science. North Holland, 1990: 365-458.

[39] Wang Y. Software engineering foundations: A software science perspective[M]. New York: Auerbach Publications, 2007.

[40] Wang Y. On the big-R notation for describing iterative and recursive behaviors[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2008, 2(1): 17-23.

[41] Wang Y. Concept algebra: A denotational mathematics for formal knowledge representation and cognitive robot learning[J]. Journal of Advanced Mathematics and Applications, 2015, 4(1): 62-87.

[42] Wang Y, Valipour M, Zatarain O A. Quantitative semantic analysis and comprehension by cognitive machine learning[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2016, 10(3): 14-28.

[43] Wang Y. Keynote: Deep reasoning and thinking beyond deep learning by cognitive robots and brain-inspired systems[C]//Proceedings of 15th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC 2016). Stanford: IEEE CS Press, 2016: 22-23.

[44] Wang Y, Widrow B, Zadeh L A, N. et al. Cognitive intelligence: Deep learning, thinking, and reasoning with brain-inspired systems[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2016, 10(4): 1-21.

[45] Sternberg R J. In search of the human mind[M]. 2nd ed. New York: Harcourt Brace & Co, 1998.

[46] Wang Y. Cognitive robots: A reference model towards intelligent authentication[J]. IEEE Robotics and Automation, 2010, 17(4): 54-62.

[47] Wang Y. Keynote: From information revolution to intelligence revolution: big data science vs. intelligence science[C]//Proceedings of 13th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC 2014). London: IEEE CS Press, 2014: 3-5.

[48] Bender E A. Mathematical methods in artificial intelligence[M]. Los Alamitos: IEEE Computer Society Press, 1997.

[49] Wang Y. On cognitive foundations and mathematical theories of knowledge science[J]. International Journal of Cognitive Informatics and Natural Intelligence, 2016, 10 (2): 1-24.

On big data algebra: A formal analytic methodology for big data science and engineering

WANG Yingxu[1,2,3], JIN Jin[2]

1. National Engineering Key Lab for Big Data System Software, School of Software, Tsinghua University, Beijing 100084, China
2. Beijing National Research Center of Information Science and Technology, Tsinghua University, Beijing 100084, China
3. International Institute of Cognitive Informatics and Cognitive Computing (ICIC), Department of Electrical and Computer Engineering, Schulich School of Engineering and Hotchkiss Brain Institute, University of Calgary, Calgary T2N1N4, Canada

Abstract Basic researches of big data science have triggered the emergence of mathematical theories of big data systems. This paper presents a rigorous analytic methodology for big data science and engineering known as Big Data Algebra (BDA). The mathematical models of big data science in BDA are formally elicited from common patterns and essences of a wide variety of big data systems. BDA reveals that any big data system is a Recursively Typed Hyperstructure (RTHS) beyond the traditional domain of pure numbers. It leads to a set of algebraic operators for big data modeling, analysis, and synthesis towards the denotational mathematical structure of BDA. The formal principles and properties of big data and their mathematical manipulations provide a theoretical framework of big data science as the basis for applications in big data engineering.

Keywords big data; mathematical model; hyper-structure; algebra; algorithms

（责任编辑　刘志远）